# CSCE 636: Deep Learning (Spring 2024) HW3 Report

1. We need to analyze the autoregressive language model and understand how the number of entries in the probability tables grows exponentially with the value of n.

The autoregressive language model, as given by equation (12.31) in the textbook "Deep Learning: Foundations and Concepts", can be represented as:

P(xn|x1,x2,...,xn-1)

Where x1,x2,...,xn-1 are the previous n-1 words in the sequence, and we want to predict the probability distribution over the next word xn.

In this model, the conditional probability P(xn|x1,x2,...,xn-1) is represented by a probability table. Each entry in this table represents the probability of observing xn given the sequence x1,x2,...,xn-1.

Let's analyze the number of entries in these probability tables as n increases:

- **a.** For the first word (x1), there are no previous words, so there is only one entry in the probability table, i.e., P(x1).
- **b.** For the second word (x2), there is one previous word (x1). Therefore, the probability table will have as many entries as there are possible values of x1. If there are V possible values for each word in the vocabulary, then the probability table for x2 will have V entries.
- **c.** For the third word (x3), there are two previous words (x1 and x2). Therefore, the probability table will have as many entries as there are possible combinations of x1 and x2, which is V^2.
- **d.** Similarly, for the nth word (xn), there are n-1 previous words. Therefore, the probability table will have as many entries as there are possible combinations of the previous n-1 words, which is V^{n-1}.

As we can see, the number of entries in the probability tables grows exponentially with the value of n, specifically as V^{n-1}, where V is the size of the vocabulary. This exponential growth occurs because each additional word in the sequence multiplies the number of possible combinations of previous words by the size of the vocabulary.

Therefore, it is evident that the number of entries in the probability tables increases rapidly as the value of n increases, making the model computationally expensive to train and use for large values of n. This exponential growth in the number of parameters is one of the challenges of scaling autoregressive language models.

- a. After the word embedding layer, the representations of words like "bank" are not inherently able to distinguish between the two occurrences in the given example sentences. This is because word embeddings typically represent words based on their context within a large corpus of text. In other words, words that appear in similar contexts are represented by similar vectors in the embedding space. In the example sentences provided, the word "bank" appears in similar contexts (as a location where one can get something), and thus, their embeddings might not capture the specific meaning or usage of "bank" in each sentence.
  For instance, in both sentences, "bank" follows the preposition "from" or "to" and is associated with the action of getting something. Therefore, the word embeddings for "bank" in both sentences might end up being quite similar since they are influenced by the context of being a destination for an action rather than capturing the semantic differences between the two types of banks (the financial institution and the side of a river).
- b. However, after applying a transformer layer on top of the word embeddings, there is a potential for the representations of "bank" to distinguish between the two occurrences in the example sentences. The transformer layer, through its self-attention mechanism, can capture more nuanced relationships between words in a sentence and encode contextual information more effectively. In the given example sentences, the transformer layer can learn to attend more to the words preceding "bank" to understand its specific usage.
  For example, in the sentence "to get cash from the bank," the transformer layer might attend more to the word "cash" and learn that "bank" here refers to a financial institution where one can withdraw or deposit money. On the other hand, in the

might attend more to the word "cash" and learn that "bank" here refers to a financial institution where one can withdraw or deposit money. On the other hand, in the sentence "to get to the other bank," the transformer layer might attend more to the words "the other" and "across the river" and learn that "bank" here refers to the side of the river. By considering the surrounding context more effectively and capturing the dependencies between words, the transformer layer has the potential to produce more meaningful representations for words like "bank," allowing it to distinguish between different occurrences based on their semantic usage in context. Therefore, while the word embeddings alone might struggle to differentiate between the two occurrences of "bank," the transformer layer enhances the model's ability to do so by leveraging contextual information more effectively.

#### 3. Coding Task

**a.** A tokenizer is used to break down a piece of text into smaller units, typically words or sub-words. It converts raw text data into a format that can be processed by machine learning models.

In the provided code, the Tokenizer class processes raw data by:

- Splitting the input text into individual tokens based on whitespace.
- Building a vocabulary of tokens encountered in the training data.
- Assigning a unique ID to each token in the vocabulary.
- Encoding input text sequences into sequences of token IDs, replacing unknown tokens with a special <unk> token ID and padding sequences to a maximum length.
- Decoding sequences of token IDs back into human-readable text.

The size of the vocabulary corresponds to the number of unique tokens encountered in the training data. In the code, the vocabulary size is dynamically determined during training and is equal to the number of unique tokens added to the vocabulary. The initial vocabulary includes four special tokens (<pad>, <s>, </s>, <unk>), and additional tokens encountered during training are incrementally added to the vocabulary. The vocabulary size is obtained using the get\_vocab\_size method of the tokenizer instance and it is 23.

- b. The maximum length of the input sequence is specified as max\_len in the command-line arguments parsed by the argparse. ArgumentParser. Its default value is set to 128, but it can be adjusted by us according to our requirements of the task and dataset.
- **c.** The self-attention mechanism involves several steps:
  - Projection: The input embeddings are projected into query, key, and value vectors. This typically involves linear transformations to project the input embeddings into higher-dimensional spaces suitable for manipulation.
  - Scoring: Each query vector is compared with all key vectors to produce attention scores. These scores represent the relevance/importance of each key with respect to the query.
  - iii. Normalization: The attention scores are normalized, usually using softmax, to ensure that the attention weights sum up to 1. This step gives weights to each value vector based on its relevance to the query.
  - iv. Weighted Sum: The normalized attention scores are used to weight the corresponding value vectors. These weighted vectors are then summed to produce the output of the attention mechanism.
  - v. Projection Back: The output from the weighted sum step is often projected back to the original embedding dimensionality through another linear transformation.

The critical step to make the self-attention mechanism causal is the masking step. Specifically, this line of code:

attn weights = attn weights.masked fill(self.mask[:, :, :L, :L] == 0, float('-inf'))

This line applies a causal mask to the attention weights, ensuring that each token can only attend to previous tokens or itself, but not to future tokens. This is crucial for models where the generation of each token depends only on the previously generated tokens during training.

The mask in the forward function of the class "GPT" is needed for a couple of reasons:

- Padding Tokens: In sequences, particularly during batch processing, sequences are padded to have uniform lengths. The mask ensures that the attention mechanism does not attend to the padded tokens, which do not contain meaningful information.
- Causal Generation: In autoregressive models like GPT, during generation, we
  don't want tokens to attend to future tokens. This would violate the
  autoregressive property. So, the mask ensures that attention is only applied
  to previous or current tokens, making the generation process causal.

#### Training process -

The training process involves iterating over multiple epochs, where each epoch consists of iterations over batches of data. Given below is a breakdown of the training process:

#### • Initialization:

- Load the dataset.
- Build the tokenizer.
- Initialize the GPT model and the trainer with the specified configurations.
- Training Loop: For each epoch,
  - Start the timer to measure the time taken for the epoch.
  - Run the training phase:
    - Set the model in training mode
    - Create a DataLoader for the training dataset.
    - Iterate over batches:
      - Move the input data to the appropriate device (CPU/GPU)
      - Perform forward pass: Generate logits and calculate the loss
      - Perform backward pass (if in training mode): Zero gradients, Scale the loss using gradient scaling (apex), Backpropagate the loss, Update model parameters, Optionally decay the learning rate based on training progress
      - Update progress and display relevant information.
  - Run the validation phase (if a validation dataset is provided):
    - Set the model in evaluation mode.
    - Create a DataLoader for the validation dataset.
    - Iterate over batches:
      - Move the input data to the appropriate device.
      - Perform forward pass to compute loss.
      - Calculate and record the average loss.
  - Record the training statistics for the epoch, such as training and validation losses.
  - Optionally save the model checkpoint if it improves over the previous best validation loss.
  - Display epoch-level training summary.

After completing all epochs, print out a results table summarizing the training process, including the architecture configurations, validation loss, and time taken per epoch.

Results: Finally, the script prints a results table summarizing the training
process, including the architecture configurations, validation loss, and time
taken per epoch.

During training, the script keeps track of the best validation loss and saves the model checkpoint if the validation loss improves. This ensures that the best model is saved for later use. Additionally, learning rate scheduling is performed based on the configured strategy (linear warm-up followed by cosine decay).

- **d.** The generation process involves using the trained GPT model to generate a sequence of actions given a set of conditions. Let us break down the generation process using a concrete example from the test dataset you provided:

  Example:
  - Conditions (IN): "turn opposite right thrice and turn opposite left"

    Expected Actions (OUT): "I\_TURN\_RIGHT I\_TURN\_RIGHT I\_TURN\_RIGHT I\_TURN\_LEFT"

    I\_TURN\_RIGHT I\_TURN\_LEFT I
  - (i) Tokenization: The input conditions ("turn opposite right thrice and turn opposite left") are tokenized using the tokenizer. This converts the input text into a sequence of tokens understandable by the model.
  - (ii) Encoding: The tokenized sequence of conditions is encoded into numerical indices using the tokenizer's vocabulary. This creates a tensor of input IDs representing the conditions.
  - (iii) Generation Loop:
    - a. The model generates the actions iteratively, token by token, starting from the encoded conditions.
    - b. At each step, the model predicts the next token in the sequence based on the previously generated tokens.
    - c. The predicted token is appended to the input sequence, and the process continues iteratively until the maximum length is reached or a termination condition is met.
  - (iv) Termination: The generation process terminates when either the maximum length is reached or a predefined stop token is generated.
  - (v) Decoding: Once the generation process is complete, the generated sequence of action tokens is decoded back into human-readable text using the tokenizer.
  - (vi) Comparison:
    - a. Finally, the generated sequence of actions is compared with the expected actions from the dataset.
    - b. If the generated actions match the expected actions, it's considered a correct prediction, and the accuracy is incremented.

In summary, the generation process involves iteratively predicting the next token in the sequence based on the previously generated tokens, until a termination condition is met. This process allows the GPT model to generate sequences of actions conditioned on the provided input conditions.

## e. Table

n_layer	n_head	n_embd	epoch	validation	time	Mean	Mean	Test
uyc.	ii_iieaa	11_611134	Сросп	loss	per	Validation	time	Accuracy
					epoch	loss	per	(Generate)
							epoch	
2	2	16	1	1.3321	8.74	0.4429	7.32	0.0662
2	2	16	2	0.8534	7.41			
2	2	16	3	0.6856	7.11			
2	2	16	4	0.6445	7.34			
2	2	16	5	0.5955	7.26			
2	2	16	6	0.5895	7.37			
2	2	16	7	0.5721	7.32			
2	2	16	8	0.5502	7.25			
2	2	16	9	0.5296	7.41			
2	2	16	10	0.5217	7.44			
2	2	16	11	0.5049	7.36			
2	2	16	12	0.5027	7.04			
2	2	16	13	0.4869	7.16			
2	2	16	14	0.4735	7.14			
2	2	16	15	0.4573	7.48			
2	2	16	16	0.4587	7.49			
2	2	16	17	0.4510	7.28			
2	2	16	18	0.4395	7.15			
2	2	16	19	0.4337	7.22			
2	2	16	20	0.4323	7.34			
2	2	16	21	0.4233	7.41			
2	2	16	22	0.4262	7.15			
2	2	16	23	0.4145	7.21			
2	2	16	24	0.4157	7.32			
2	2	16	25	0.4054	7.16			
2	2	16	26	0.4069	7.23			
2	2	16	27	0.4012	7.18			
2	2	16	28	0.3985	7.05			
2	2	16	29	0.3921	7.31			
2	2	16	30	0.3909	7.13			
2	2	16	31	0.3854	7.20			
2	2	16	32	0.3890	7.36			
2	2	16	33	0.3883	7.33			
2	2	16	34	0.3861	7.37			
2	2	16	35	0.3820	7.36			
2	2	16	36	0.3742	7.30			
2	2	16	37	0.3754	7.26			
2	2	16	38	0.3787	7.24			
2	2	16	39	0.3705	7.43			
2	2	16	40	0.3682	7.44			
2	2	16	41	0.3703	7.17			

	۔ ا	l	1 _	_	_ 1		Ì	
2	2	16	42	0.3680	7.27			
2	2	16	43	0.3655	6.97			
2	2	16	44	0.3614	7.33			
2	2	16	45	0.3628	7.14			
2	2	16	46	0.3647	7.42			
2	2	16	47	0.3602	7.46			
2	2	16	48	0.3582	7.68			
2	2	16	49	0.3585	7.34			
2	2	16	50	0.3574	7.32			
2	2	16	51	0.3592	6.99			
2	2	16	52	0.3524	7.33			
2	2	16	53	0.3532	7.29			
2	2	16	54	0.3579	7.37			
2	2	16	55	0.3545	7.50			
2	2	16	56	0.3563	7.44			
2	2	16	57	0.3585	7.33			
2	2	16	58	0.3538	7.40			
2	2	16	59	0.3585	7.31			
2	2	16	60	0.3551	7.30			
2	4	16	1	1.3210	8.21	0.3544	7.15	0.1698
2	4	16	2	0.8383	7.18			
2	4	16	3	0.6807	7.12			
2	4	16	4	0.6174	7.04			
2	4	16	5	0.5637	7.24			
2	4	16	6	0.5409	7.06			
2	4	16	7	0.5131	6.98			
2	4	16	8	0.4971	7.20			
2	4	16	9	0.4764	7.12			
2	4	16	10	0.4656	7.11			
2	4	16	11	0.4542	7.22			
2	4	16	12	0.4456	7.03			
2	4	16	13	0.4273	7.13			
2	4	16	14	0.4185	7.20			
2	4	16	15	0.4099	7.32			
2	4	16	16	0.3958	7.01			
2	4	16	17	0.3909	7.13			
2	4	16	18	0.3816	7.01			
2	4	16	19	0.3732	7.11			
2	4	16	20	0.3609	7.02			
2	4	16	21	0.3572	7.28			
2	4	16	22	0.3515	7.14			
2	4	16	23	0.3356	7.10			
2	4	16	24	0.3346	6.97			
2	4	16	25	0.3265	6.86			
2	4	16	26	0.3111	7.10			
2	4	16	27	0.3050	7.36			

l a	۱ ۵	1 40	۱ ۵۵	0.0075	7.05			
2	4	16	28	0.2975	7.05			
2	4	16	29	0.2975	7.11			
		16	30	0.2959	6.93			
2	4	16	31	0.2994	7.15			
2	4	16	32	0.2790	7.29			
2	4	16	33	0.2731	7.20			
2	4	16	34	0.2700	6.99			
2	4	16	35	0.2668	7.32			
2	4	16	36	0.2594	7.24			
2	4	16	37	0.2604	7.04			
2		16	38	0.2593	7.01			
2	4	16 16	39	0.2544	7.25			
2	4		40	0.2537	7.12			
2	4	16	41	0.2469	7.25			
2	4	16	42	0.2479	7.25			
2	4	16	43	0.2456	7.22			
2	4	16 16	44	0.2467	7.26			
2	4	16	45	0.2397	7.10			
2	4		46	0.2419	7.06			
2	4	16 16	47	0.2441	7.08			
2	4		48	0.2411	7.08			
2	4	16	49	0.2373	7.16			
2		16	50	0.2381	7.06			
2	4	16	51	0.2411	6.99			
2	4	16 16	52	0.2371	7.12			
2	4	16	53 54	0.2369	7.19 7.23			
2	4	16	55	0.2377 0.2350	7.23			
2	4	16	56	0.2378	7.21			
2	4	16	57	0.2378				
2	4	16	58	0.2422	7.07 7.12			
2	4	16	59	0.2378	7.12			
2	4	16	60	0.2378	7.20			
4	4	32	1	0.2348	11.48	0.1000	10.25	0.9922
4	4	32	2	0.5521	10.38	0.1000	10.23	0.5522
4	4	32	3	0.4576	10.15			
4	4	32	4	0.3655	10.17			
4	4	32	5	0.3162	10.40			
4	4	32	6	0.2680	10.18			
4	4	32	7	0.2353	10.37			
4	4	32	8	0.2098	10.21			
4	4	32	9	0.1876	10.21			
4	4	32	10	0.1745	10.24			
4	4	32	11	0.1463	10.22			
4	4	32	12	0.1332	10.20			
4	4	32	13	0.1151	10.16			
				0.1101	-0.10	<u> </u>	<u> </u>	

	j .	<b>j</b> 1	Ī	] I	
4	4	32	14	0.1044	10.22
4	4	32	15	0.0885	10.32
4	4	32	16	0.0888	10.24
4	4	32	17	0.0769	10.34
4	4	32	18	0.0740	10.21
4	4	32	19	0.0711	10.20
4	4	32	20	0.0684	10.34
4	4	32	21	0.0995	10.26
4	4	32	22	0.0680	10.22
4	4	32	23	0.0746	10.27
4	4	32	24	0.0484	10.15
4	4	32	25	0.0826	10.16
4	4	32	26	0.0524	10.23
4	4	32	27	0.0530	10.18
4	4	32	28	0.0448	10.20
4	4	32	29	0.0364	10.14
<u>.</u> 4	4	32	30	0.0408	10.39
4	4	32	31	0.0431	10.41
4	4	32	32	0.0325	10.22
4	4	32	33	0.0516	10.24
4	4	32	34	0.0310	10.20
4	4	32	35	0.0444	10.18
<u>.</u> 4	4	32	36	0.0351	10.24
<u>.</u> 4	4	32	37	0.0416	10.28
4	4	32	38	0.0378	10.23
4	4	32	39	0.0362	10.15
4	4	32	40	0.0330	10.40
4	4	32	41	0.0320	10.36
4	4	32	42	0.0266	10.21
<del>-</del> _4	4	32	43	0.0250	10.21
<del>_</del> _4	4	32	44	0.0291	10.18
4	4	32	45	0.0251	10.13
4	4	32	46	0.0233	10.23
<del>-</del>	4	32	47	0.0277	10.11
4	4	32	48	0.0233	10.11
4	4	32	49	0.0262	10.21
4	4	32	50	0.0202	10.16
4	4	32	51	0.0243	10.19
4	4	32	52	0.0264	10.19
4	4	32	53	0.0262	10.10
4	4	32	54	0.0262	10.28
4	4	32			
4	4	32	55 56	0.0207 0.0254	10.17 10.20
4	4				
		32	57	0.0236	10.34
4	4	32 32	58 59	0.0322 0.0218	10.14
4	4	32	צכ	0.0218	10.30

4	4	32	60	0.0260	10.16			
6	8	64	1	0.5135	15.27	0.0432	14.49	0.9931
6	8	64	2	0.3070	13.81			
6	8	64	3	0.2427	14.17			
6	8	64	4	0.1866	14.35			
6	8	64	5	0.1432	14.50			
6	8	64	6	0.1127	14.55			
6	8	64	7	0.0858	14.66			
6	8	64	8	0.1301	14.64			
6	8	64	9	0.0524	14.44			
6	8	64	10	0.0673	14.63			
6	8	64	11	0.0539	14.49			
6	8	64	12	0.0458	14.62			
6	8	64	13	0.0420	14.42			
6	8	64	14	0.0357	14.53			
6	8	64	15	0.0347	14.39			
6	8	64	16	0.0279	14.55			
6	8	64	17	0.0238	14.54			
6	8	64	18	0.0291	14.68			
6	8	64	19	0.0278	14.50			
6	8	64	20	0.0199	14.51			
6	8	64	21	0.0304	14.24			
6	8	64	22	0.0260	14.46			
6	8	64	23	0.0262	14.67			
6	8	64	24	0.0209	14.09			
6	8	64	25	0.0228	14.27			
6	8	64	26	0.0144	14.17			
6	8	64	27	0.0130	14.51			
6	8	64	28	0.0132	14.29			
6	8	64	29	0.0130	14.51			
6	8	64	30	0.0117	14.50			
6	8	64	31	0.0163	14.73			
6	8	64	32	0.0150	14.49			
6	8	64	33	0.0081	14.59			
6	8	64	34	0.0082	14.75			
6	8	64	35	0.0108	14.57			
6	8	64	36	0.0210	14.65			
6	8	64	37	0.0062	14.64			
6	8	64	38	0.0119	14.69			
6	8	64	39	0.0111	14.68			
6	8	64	40	0.0071	14.17			
6	8	64	41	0.0074	14.41			
6	8	64	42	0.0101	14.35			
6	8	64	43	0.0035	14.34			
6	8	64	44	0.0093	14.32			
6	8	64	45	0.0129	14.36			

	i	ı	ı	1	1		Ī	1 1
6	8	64	46	0.0027	14.66			
6	8	64	47	0.0046	14.32			
6	8	64	48	0.0064	14.70			
6	8	64	49	0.0051	14.47			
6	8	64	50	0.0045	14.61			
6	8	64	51	0.0042	14.53			
6	8	64	52	0.0022	14.43			
6	8	64	53	0.0034	14.55			
6	8	64	54	0.0060	14.56			
6	8	64	55	0.0054	14.52			
6	8	64	56	0.0032	14.57			
6	8	64	57	0.0053	14.52			
6	8	64	58	0.0024	14.56			
6	8	64	59	0.0025	14.38			
6	8	64	60	0.0017	14.38			
8	8	128	1	0.4003	18.47	0.0257	17.02	1.0000
8	8	128	2	0.2010	16.97			
8	8	128	3	0.1376	17.19			
8	8	128	4	0.1069	16.81			
8	8	128	5	0.1006	16.93			
8	8	128	6	0.0845	17.13			
8	8	128	7	0.0629	16.98			
8	8	128	8	0.0692	17.09			
8	8	128	9	0.0350	16.83			
8	8	128	10	0.0497	17.09			
8	8	128	11	0.0297	16.78			
8	8	128	12	0.0314	16.96			
8	8	128	13	0.0376	16.81			
8	8	128	14	0.0168	17.08			
8	8	128	15	0.0186	16.86			
8	8	128	16	0.0146	16.68			
8	8	128	17	0.0157	16.86			
8	8	128	18	0.0152	17.12			
8	8	128	19	0.0225	17.01			
8	8	128	20	0.0074	17.37			
8	8	128	21	0.0064	16.90			
8	8	128	22	0.0082	16.81			
8	8	128	23	0.0068	16.86			
8	8	128	24	0.0037	16.84			
8	8	128	25	0.0106	16.92			
8	8	128	26	0.0040	16.89			
8	8	128	27	0.0029	16.90			
8	8	128	28	0.0035	17.06			
8	8	128	29	0.0011	17.11			
8	8	128	30	0.0019	16.82			
8	8	128	31	0.0032	16.90			

	۰ -	400	22	00000	46 77		İ	]
8	8	128	32	0.0009	16.77			
8	8	128	33	0.0018	17.10			
8	8	128	34	0.0037	16.96			
8	8	128	35	0.0074	16.90			
8	8	128	36	0.0067	17.07			
8	8	128	37	0.0044	17.47			
8	8	128	38	0.0005	17.06			
8	8	128	39	0.0016	17.17			
8	8	128	40	0.0005	17.11			
8	8	128	41	0.0007	17.07			
8	8	128	42	0.0001	16.94			
8	8	128	43	0.0002	17.07			
8	8	128	44	0.0014	16.82			
8	8	128	45	0.0001	17.00			
8	8	128	46	0.0001	17.28			
8	8	128	47	0.0003	17.12			
8	8	128	48	0.0004	16.77			
8	8	128	49	0.0001	16.97			
8	8	128	50	0.0000	16.94			
8	8	128	51	0.0008	17.05			
8	8	128	52	0.0001	16.98			
8	8	128	53	0.0001	16.94			
8	8	128	54	0.0001	17.01			
8	8	128	55	0.0010	17.46			
8	8	128	56	0.0000	17.17			
8	8	128	57	0.0001	16.82			
8	8	128	58	0.0000	16.93			
8	8	128	59	0.0001	17.07			
8	8	128	60	0.0000	17.08			
10	16	256	1	0.3069	21.43	0.0170	20.64	1.0000
10	16	256	2	0.1537	19.80			
10	16	256	3	0.1245	20.25			
10	16	256	4	0.0822	20.18			
10	16	256	5	0.0406	20.24			
10	16	256	6	0.0389	20.09			
10	16	256	7	0.0594	20.68			
10	16	256	8	0.0229	20.69			
10	16	256	9	0.0234	20.64			
10	16	256	10	0.0277	20.96			
10	16	256	11	0.0146	20.61			
10	16	256	12	0.0242	20.63			
10	16	256	13	0.0107	20.57			
10	16	256	14	0.0116	20.46			
10	16	256	15	0.0136	20.57			
10	16	256	16	0.0068	20.64			
10	16	256	17	0.0051	20.58			

10	16	256	18	0.0055	20.84
10	16	256	19	0.0056	20.57
10	16	256	20	0.0052	20.38
10	16	256	21	0.0052	21.00
10	16	256	22	0.0033	20.70
10	16	256	23	0.0025	20.72
10	16	256	24	0.0021	20.69
10	16	256	25	0.0025	21.17
10	16	256	26	0.0015	20.77
10	16	256	27	0.0011	20.86
10	16	256	28	0.0020	20.80
10	16	256	29	0.0009	20.62
10	16	256	30	0.0009	20.38
10	16	256	31	0.0016	20.47
10	16	256	32	0.0006	21.02
10	16	256	33	0.0015	20.24
10	16	256	34	0.0009	20.87
10	16	256	35	0.0028	20.66
10	16	256	36	0.0005	21.09
10	16	256	37	0.0002	21.14
10	16	256	38	0.0003	20.68
10	16	256	39	0.0009	20.64
10	16	256	40	0.0003	20.52
10	16	256	41	0.0003	20.87
10	16	256	42	0.0002	20.96
10	16	256	43	0.0003	20.70
10	16	256	44	0.0004	20.61
10	16	256	45	0.0002	20.77
10	16	256	46	0.0002	20.69
10	16	256	47	0.0002	20.46
10	16	256	48	0.0004	20.98
10	16	256	49	0.0003	20.69
10	16	256	50	0.0003	20.40
10	16	256	51	0.0003	20.38
10	16	256	52	0.0000	20.60
10	16	256	53	0.0002	20.91
10	16	256	54	0.0002	20.52
10	16	256	55	0.0001	20.80
10	16	256	56	0.0002	20.49
10	16	256	57	0.0001	20.32
10	16	256	58	0.0004	20.35
10	16	256	59	0.0003	20.55
10	16	256	60	0.0002	20.48

Based on the provided data, we can observe the impact of the hyperparameters (number of layers, number of heads, and embedding size) on the model performance:

- Number of Layers (n\_layer): As the number of layers increases, the model
  tends to perform better in terms of mean validation loss and test accuracy.
  This is evident from the trend where the mean validation loss decreases as
  the number of layers increases. However, increasing the number of layers
  also leads to an increase in the time per epoch, as more computations are
  required for training.
- Number of Heads (n\_head): Increasing the number of heads generally leads to improved performance, as seen in the decreasing trend of mean validation loss and increasing trend of test accuracy. However, the improvement is not as pronounced as increasing the number of layers.
- Embedding Size (n\_embd): Larger embedding sizes tend to result in better
  performance, as indicated by the decreasing trend in mean validation loss
  and increasing trend in test accuracy. However, larger embedding sizes also
  require more computational resources and may increase training time.

So, increasing the number of layers, heads, and embedding size generally improves model performance in terms of validation loss and test accuracy. However, this improvement comes at the cost of increased computational complexity and training time. Therefore, there is a trade-off between model performance and computational resources.

f. Used the data split 'length'. The split is designed to evaluate the model's ability to generalize and produce accurate outputs for longer action sequences that contain combinations of familiar elements (verbs, modifiers, and conjunctions) seen during training but in novel combinations or lengths. This evaluation assesses the model's capacity to effectively combine and extend learned actions and linguistic structures to generate coherent and contextually appropriate sequences, even when faced with previously unseen combinations or lengths of actions.

n_layer	n_head	n_embd	validation loss	time per epoch	mean validation loss	mean time per epoch	Testing Accuracy (Generate)
50	2	16	1.3816	97.57	0.1599	85.75	0.00
50	2	16	1.0269	86.88			
50	2	16	0.8456	86.89			
50	2	16	0.7739	87.82			
50	2	16	0.7046	87.01			
50	2	16	0.6337	87.13			
50	2	16	0.5708	87.60			
50	2	16	0.4480	86.94			
50	2	16	0.3884	87.45			

50	2	16	0.3069	87.30
50	2	16	0.2482	87.28
50	2	16	0.2058	87.34
50	2	16	0.1696	87.22
50	2	16	0.1351	87.16
50	2	16	0.1160	87.07
50	2	16	0.1353	86.49
50	2	16	0.0813	86.01
50	2	16	0.1120	87.33
50	2	16	0.0697	87.22
50	2	16	0.0635	87.53
50	2	16	0.1273	86.09
50	2	16	0.0633	86.77
50	2	16	0.0598	83.87
50	2	16	0.0471	83.62
50	2	16	0.0432	83.41
50	2	16	0.0627	86.62
50	2	16	0.0622	85.15
50	2	16	0.0428	86.73
50	2	16	0.0350	85.15
50	2	16	0.0313	85.83
50	2	16	0.0344	84.07
50	2	16	0.0369	85.96
50	2	16	0.0296	86.35
50	2	16	0.0283	85.82
50	2	16	0.0478	86.16
50	2	16	0.0235	85.60
50	2	16	0.0353	82.26
50	2	16	0.0209	82.60
50	2	16	0.0203	83.15
50	2	16	0.0211	85.60
50	2	16	0.0189	85.44
50	2	16	0.0173	
				86.89
50	2	16	0.0171	85.25
50	2	16	0.0163	85.59
50	2	16	0.0159	87.78
50	2	16	0.0161	91.30
50	2	16	0.0151	86.03
50	2	16	0.0155	85.81
50	2	16	0.0172	85.84
50	2	16	0.0151	86.19
50	2	16	0.0132	85.80
50	2	16	0.0181	82.20
50	2	16	0.0137	82.28
50	2	16	0.0151	82.62

50	2	16	0.0146	81.54
50	2	16	0.0129	82.08
50	2	16	0.0147	81.85
50	2	16	0.0139	81.74
50	2	16	0.0129	82.06
50	2	16	0.0138	82.79

### Insights:

- Our CSA based GPT model, by design, imposes causality constraints on the attention mechanism, ensuring that each token attends only to previous tokens or itself. This is particularly useful for autoregressive tasks where the generation of each token depends only on previously generated tokens. On the other hand, the attention mechanism used in the GRU model might be more flexible and adaptive, allowing the model to attend to any part of the input sequence based on its relevance to the current context. This flexibility could be advantageous, especially for tasks with longer input sequences.
- Our model, offers high model capacity with its multiple layers, however, the ability of CSA to generalize to longer sequences might be limited compared to models equipped with more flexible attention mechanisms. The GRU model mentioned in the paper, despite having a smaller capacity, achieves competitive performance by effectively leveraging attention mechanisms, indicating the importance of attention in handling complex sequence modelling tasks.

## **Appendix**

'Simple' Data Split - Test Accuracies

```
[apurva.mandalika@g062 code]$ python main.py --task generate --n_layer 2 --n_head 2 --n_embd 16
The file './tokenizer/simple_vocab.json' exists. Loading tokenizer.
{'cpad': 0, 'cs>': 1, 'c/s>': 2, 'cunk': 3, 'I_URN_RIGHT': 4, 'I_UMP': 5, 'I_WALK': 6, 'I_TURN_LEFT': 7, 'I_RUN': 8, 'I_LOOK': 9, 'jump': 10, 'opposite': 11, 'right': 12, 'twice': 13, 'and': 14, 'turn': 15, 'thrice': 16, 'run': 17, 'left': 18, 'after': 19, 'walk': 20, 'around': 21, 'look': 22}
loading model
total params: 9408
Accuracy: 0.0662367505978001
[apurva.mandalika@g062 code]$

[apurva.mandalika@g062 code]$

[apurva.mandalika@g062 code]$

[apurva.mandalika@g062 code]$

[apurva.mandalika@g062 code]$
                                                                                                                                                                                                                                                                                                                                                                                                 | 4182/4182 [13:09<00:00, 5.29it/s]
   [apurva.mandalika@g062 code]$ python main.py --task generate --n_layer 2 --n_head 4 --n_embd 16
The file './tokenizer/simple_vocab.json' exists. Loading tokenizer.
{'<pad>': 0, '<s>': 1, '</s>': 2, '<unk>': 3, 'I_URN_RIGHT': 4, 'I_JUMP': 5, 'I_WALK': 6, 'I_TURN_LEFT': 7, 'I_RUN': 8, 'I_LOOK': 9, 'jump': 10, 'opposite': 11, 'right': 12, 'twice'
: 13, 'and': 14, 'turn': 15, 'thrice': 16, 'run': 17, 'left': 18, 'after': 19, 'walk': 20, 'around': 21, 'look': 22}
 : 13, and 14, chilling index loading model total params: 9408 Accuracy: 0.1698: 100% | Test accuracy: 0.1697/522/16403634 | [apurva.mandallka@g062 code]$
                                                                                                                                                                                                                                                                                                                                                                                               | 4182/4182 [13:15<00:00, 5.26it/s]
 [apurva.mandalika@g080 code]$ python main.py --task generate --n_layer 4 --n_embd 32

The file './tokenizer/simple_vocab.json' exists. Loading tokenizer.

{'<pad>': 0, '<s>': 1, '</s>': 2, '<unk>': 3, 'I_TURN_RIGHT': 4, 'I_JUMP': 5, 'I_WALK': 6, 'I_TURN_LEFT': 7, 'I_RUN': 8, 'I_LOOK': 9, 'jump': 10, 'opposite': 11, 'right': 12, 'twice': 13, 'and': 14, 'turn': 15, 'thrice': 16, 'run': 17, 'left': 18, 'after': 19, 'walk': 20, 'around': 21, 'look': 22}

loading model

total params: 56512

Accuracy: 0.9122: 100%|

Test accuracy: 0.9122: 100%|

Test accuracy: 0.9122:29459588713

[apurva.mandalika@g080 code]$

[apurva.mandalika@g080 code]$

[apurva.mandalika@g080 code]$ python main.py --task generate --n layer 6 --n head 8 --n embd 64
 [apurva.mandalika@g@00 code]$ python main.py --task generate --n_layer 6 --n_head 8 --n_embd 64
The file './tokenizer/simple_vocab.json' exists. Loading tokenizer.
{'qad>' o, 'cs>' l, '/s>' 2, 'dunk>' 3, 'I_UNN_RIGHT': 4, 'I_JUMP': 5, 'I_MALK': 6, 'I_TUNN_LEFT': 7, 'I_RUN': 8, 'I_LOOK': 9, 'jump': 10, 'opposite': 11, 'right': 12, 'twice'
: 13, 'and': 14, 'turn': 15, 'thrice': 16, 'run': 17, 'left': 18, 'after': 19, 'walk': 20, 'around': 21, 'look': 22}
    oading model
  loading model
total params: 311296
Accuracy: 0.9931: 100%|
Test accuracy: 0.993065518890483
[apurva.mandalika@g080 code]$ []
                                                                                                                                                                                                                                                                                                                                                                                                | 4182/4182 [35:04<00:00, 1.99it/s]
[apurva.mandalika@g009 code]$ python main.py --task generate --n_layer 10 --n_head 16 --n_embd 256
The file '-/tokenizer/simple_vocab.json' exists. Loading tokenizer.
{'<pad>': 0, '<s>': 1, '</s>': 2, '<unk>': 3, 'I_TURN_RIGHT': 4, 'I_JUMP': 5, 'I_WALK': 6, 'I_TURN_LEFT': 7, 'I_RUN': 8, 'I_LOOK': 9, 'jump': 10, 'opposite': 11, 'right': 12, 'twice'
: 13, 'and': 14, 'turn': 15, 'thrice': 16, 'run': 17, 'left': 18, 'after': 19, 'walk': 20, 'around': 21, 'look': 22}
| coditer metal
 : 13, 'and': 14, 'turn'
loading model
total panams: 7943168
Accuracy: 1.0000: 39%
  Accuracy: 1.0000:
Accuracy: 1.0000:
Accuracy: 1.0000:
    Accuracy: 1.0000: 100%
                                                                                                                                                                                                                                                                                                                                                                                                     4182/4182 [56:08<00:00, 1.24it/s]
        est accuracy: 1.0
nurva.mandalika@g009 code]$
```

'length' Data Split – Train loss, Validation loss and Test Accuracies

```
09 code]$ python main.py --task train --data_split 'length' --n_layer 50
        [apurva.mandalika@g009 code]$ python main.py --task
Downloading data: 100%|
Downloading tata: 100%|
Generating train split: 100%|
Generating test split: 100%|
Building tokenizer at ./tokenizer/length_vocab.json.
Building tokenizer for actions: 100%|
Building tokenizer for commands: 100%|
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              | 450k/450k [00:00<00:00, 2.88MB/s] | 123k/123k [00:00<00:00, 1.92MB/s] | 123k/123k [00:00<00:00, 1.92MB/s] | 16990/16990 [00:00<00:00, 185667.49 examples/s] | 3920/3920 [00:00<00:00, 388793.11 examples/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             16990/16990 [00:00<00:00, 28534.60it/s]
16990/16990 [00:00<00:00, 31311.26it/s]
        | 10990/10990 [00:00000:00, 31311.01175] tokenizer saved {\checkizer saved sav
            loading model
        loading model
total params: 166848
epoch 1 iter 477: train loss 1.42470. lr 3.9978e-04: 100%
test loss: %f 1.3816234270731609
epoch_valid_loss: 1.3816234270731609, epoch_train_loss: 2.019174261571972, epoch: 1
Saving at epoch 1: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt
step_train_loss: 1.3530999422073364 train_step: 500, learning_rate: 0.0003997570245178301
epoch 2 iter 477: train loss 0.98469. lr 3.9902e-04: 100%
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    478/478 [01:33<00:00, 5.09it/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 22/478 [00:03<01:19, 5.72it/s]
478/478 [01:23<00:00, 5.71it/s]
            test loss: %f 1.0268689802399389
      | 44/478 [00:07<01:15, 5.71it/s]
| 478/478 [01:23<00:00, 5.71it/s]
     test loss: %f 0.8456329692293996 epoch_train_loss: 0.9815252440743866, epoch: 3 Saving at epoch 3: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.9417285323143905 train_step: 1500, learning_rate: 0.0003975036175640473 epoch 4 iter 477: train loss 0.79178. lr 3.9590e-04: 100%| test loss: %f 0.7738873450844376 epoch_valid_loss: 0.7738873450844376, epoch_train_loss: 0.8806835711686681, epoch: 4 Saving at epoch 4: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.8316246271133423 train_step: 2000, learning_rate: 0.0003954972234635432 epoch 5 iter 477: train loss 0.71059. lr 3.9354e-04: 100%| test loss: %f 0.7046002171657704 epoch_train_loss: 0.8155013565987224, epoch: 5
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 66/478 [00:11<01:12, 5.66it/s]
| 478/478 [01:24<00:00, 5.64it/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 88/478 [00:15<01:07, 5.75it/s]
| 478/478 [01:23<00:00, 5.70it/s]
epoch_valid_loss: 0.7046002171657704, epoch_train_loss: 0.8155013565987224, epoch: 5

Saving at epoch 5: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt
step_train_loss: 0.775398313999176 train_step: 2500, learning_rate: 0.0003929117158854165
epoch 6 iter 477: train loss 0.779900. lr 3.9065e-04: 100%|
test loss: % f 0.6336997482511733
epoch_valid_loss: 0.6336997482511733, epoch_train_loss: 0.7570871221969317, epoch: 6
Saving at epoch 5: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt
step_train_loss: 0.7298206686973572 train_step: 3000, learning_rate: 0.0003897547537736588
epoch 7 iter 477: train loss: 0.67298. lr 3.8725e-04: 100%|
test loss: % f 0.5708381583293279
epoch_valid_loss: 0.5708381583293279, epoch_train_loss: 0.7047941900446824, epoch: 7
Saving at epoch 7: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt
step_train_loss: 0.5971279144287109 train_step: 3500, learning_rate: 0.00038603568886879546
epoch 8 iter 477: train loss 0.65885. lr 3.8334e-04: 100%|
test loss: % f 0.4480134127316652
epoch_valid_loss: 0.4480134127316652
epoch_valid_loss: 0.4480134127316652
epoch_valid_loss: 0.6112558245658875 train_step: 4000, learning_rate: 0.000381765538005605
epoch 9 iter 477: train loss 0.43980. ln 3.78394e-04: 100%|
test loss: % f 0.38839665331222395
epoch_valid_loss: 0.597471997503457, epoch: 9
Saving at epoch 9.
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 110/478 [00:19<01:04, 5.66it/s]
| 478/478 [01:24<00:00, 5.69it/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 132/478 [00:23<01:00, 5.70it/s]
| 478/478 [01:24<00:00, 5.66it/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         154/478 [00:27<00:56, 5.69it/s]
478/478 [01:23<00:00, 5.70it/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 176/478 [00:30<00:53, 5.63it/s]
| 478/478 [01:24<00:00, 5.67it/s]
 epoch 9 1ter 47: train 10ss 0.33940. Ir 3.7894e-04: 100%|
test 10ss: % 6 0.38839665331222395 epoch_train_loss: 0.5297471997503457, epoch: 9
epoch_valid_loss: 0.38839665331222395, epoch_train_loss: 0.5297471997503457, epoch: 9
Saving at epoch 9: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt
step_train_loss: 0.45415839552879333 train_step: 4500, learning_rate: 0.00037695695047839723
epoch 10 iter 477: train loss 0.38224. lr 3.7405e-04: 100%|
test loss: % 6 0.30689116374210074, epoch_train_loss: 0.4657395201497497, epoch: 10
Saving at epoch 10: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt
step_train_loss: 0.376885398961404004 train_step: 5000, learning_rate: 0.00037162416781880804
epoch 11 iter 477: train loss 0.33641. lr 3.6869e-04: 100%|
test loss: % 6 0.24817123264074326
epoch_valid_loss: 0.24817123264074326, epoch_train_loss: 0.4155208083126834, epoch: 11
Saving at epoch 11: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt
step_train_loss: 0.3754410147666931 train_step: 5500, learning_rate: 0.0003657829893634123
epoch 12 iter 477: train loss 0.35893. lr 3.6287e-04: 100%|
test loss: % 6 0.20577994264938212
epoch_valid_loss: 0.20577994264938212, epoch_train_loss: 0.3710573667886367, epoch: 12
Saving at epoch 12: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt
step_train_loss: 0.3281327486693828 train_step: 6000, learning_rate: 0.0003594507246649326
epoch_valid_loss: 0.20577994264938212
epoch_valid_loss: 0.20577994264938
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 198/478 [00:34<00:48, 5.73it/s]
| 478/478 [01:24<00:00, 5.68it/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 220/478 [00:38<00:45, 5.72it/s]
| 478/478 [01:24<00:00, 5.68it/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         242/478 [00:42<00:41, 5.69it/s]
478/478 [01:24<00:00, 5.68it/s]
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | 264/478 [00:46<00:37, 5.71it/s]
| 478/478 [01:24<00:00, 5.69it/s]
```

epoch 15 iter 477: train loss 0.27436. lr 3.4284e-04: 100%	478/478 [01:23<00:00, 5.69it/s]
test loss: Xf 0.11595677950750103	
epoch_valid_loss: 0.11595657950750103, epoch_train_loss: 0.2738531071334204, epoch: 15 Saving at epoch 15: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	
Saving at epocn is: ./cono_gpt/weignts/Mone_lengtnspilt_skiayer_zhead_leemad_szos.pt step_train_loss: 0.25593548607826233 train_step: 7500, learning_rate: 0.00033770187882238545	330/478 [00:58<00:25, 5.74it/s]
step_train_sss. 0.2239334000/020230 train_step. /300, tead fining_rate. 0.000397/01070022303450 epoch 16 iter 477: train loss 0.20920. lr 3.5536e-04: 100%	478/478 [01:23<00:00, 5.73it/s]
epoch to ite* 47. train tos 6.20520. ir 3.33300-04. 1004   test loss: % 0.1352660658734816	[ 4/8/4/8 [01.23(00.00, 3./31(/5]
epoch_valid_loss: 0.1352660658734816, epoch_train_loss: 0.2595901943542868, epoch: 16	
step_train_loss: 0.17839477956295613 train_step: 8000, learning_rate: 0.00022960651359957967	352/478 [01:00<00:22, 5.70it/s]
epoch 17 iter 477: train loss 0.25633. lr 3.2752e-04: 100%	478/478 [01:22<00:00, 5.77it/s]
test loss: %f 0.08133871956831878	,,, ,, ,, ,, ,, ,
epoch valid loss: 0.08133871956831878, epoch_train_loss: 0.22959262999281224, epoch: 17	
Saving at epoch 17: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	
step_train_loss: 0.21427012979984283 train_step: 8500, learning_rate: 0.0003211272206482685	374/478 [01:06<00:18, 5.70it/s]
epoch 18 iter 477: train loss 0.19505. lr 3.1934e-04: 100%	478/478 [01:24<00:00, 5.67it/s]
test loss: %f 0.11203097965982226	
epoch_valid_loss: 0.11203097965982226, epoch_train_loss: 0.21735326581046172, epoch: 18	
step_train_loss: 0.20009490847587585 train_step: 9000, learning_rate: 0.00031228911780853354	396/478 [01:09<00:14, 5.72it/s]
epoch 19 iter 477: train loss 0.21242. lr 3.1083e-04: 100%	478/478 [01:24<00:00, 5.68it/s]
test loss: %f 0.06970568711834925	
epoch_valid_loss: 0.06970568711834925, epoch_train_loss: 0.1983176127200845, epoch: 19	
Saving at epoch 19: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	
step_train_loss: 0.2263641357421875 train_step: 9500, learning_rate: 0.0003031183674392653	418/478 [01:13<00:10, 5.74it/s]
epoch 20 iter 477: train loss 0.13543. lr 3.0202e-04: 100%	478/478 [01:24<00:00, 5.66it/s]
test loss: %f 0.06352007120019859	
epoch_valid_loss: 0.06352007120019859, epoch_train_loss: 0.1882161243378368, epoch: 20	
Saving at epoch 20: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	
step_train_loss: 0.12573248147964478 train_step: 10000, learning_rate: 0.0002936421717623602	440/478 [01:16<00:06, 6.05it/s]
epoch 21 iter 477: train loss 0.13359. lr 2.9293e-04: 100%	478/478 [01:22<00:00, 5.76it/s]
test loss: %f 0.12730575766828325	
epoch_valid_loss: 0.12730575766828325, epoch_train_loss: 0.1779476337010142, epoch: 21	
step_train_loss: 0.13674236834049225 train_step: 10500, learning_rate: 0.0002838885839598702	462/478 [01:21<00:02, 6.03it/s]
epoch 22 iter 477: train loss 0.16842. lr 2.8359e-04: 100%	478/478 [01:23<00:00, 5.71it/s]
test loss: %f 0.06330315358246918	
epoch_valid_loss: 0.06330315358246918, epoch_train_loss: 0.16531385019982708, epoch: 22	
Saving at epoch 22: /cond_gpt/weights/None_lengthsplit_59layer_2head_16embd_32bs.pt	1 470 470 504 20 00 00 5 00:44 3
epoch 23 iter 477: train loss 0.13672. lr 2.7403e-04: 100%	478/478 [01:20<00:00, 5.92it/s]
test loss: %f 0.05981088084755121	
step_train_loss: 0.1299994989299774 train_step: 11500, learning_rate: 0.0002636687494788558	28/478 [00:04<01:15, 5.97it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%	28/478 [00:04<01:15, 5.97it/s]   478/478 [01:20<00:00, 5.95it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923	
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25	
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250803923 epoch_valid_loss: 0.04317793250803923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt	478/478 [01:20<00:00, 5.95it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: \$\forall 6.0817793250808392\$ epoch_valid_loss: 0.04317793250803923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch_25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1337837494818802 train_step: 12000, learning_rate: 0.00025352952516845316	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%	478/478 [01:20<00:00, 5.95it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: Xf 0.043177932508033923 epoch_valid_loss: 0.043177932508033923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.135783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: Xf 0.06270570887459649	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: Xf 0.00317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.13357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: Xf 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32Ds.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.111458584666625214 train_step: 12500, learning_rate: 0.00024269198632935544	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.135783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_valld_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837949818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch_27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: %f 0.062707788773842047762	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32Ds.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.062705708887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.111458584666625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch_27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: %f 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: \$\frac{\phi}{1}.0817793250083923} = \text{poch_valid_loss: 0.04317793250083923}, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: \$\frac{\phi}{1}.06270570887459649} epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145853466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.16264. lr 2.3404e-04: 100%  test loss: \$\frac{\phi}{1}.08202773842047762} epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.08220773842047762 epoch_valid_loss: 0.08220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.08284437084197998 train_step: 13000, learning_rate: 0.00023199823520207903	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_valld_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: %f 0.062706773842047762 epoch_valid_loss: 0.062706773842047762 epoch_valid_loss: 0.062206773842047762 epoch_valid_loss: 0.0622067738420467762 epoch_valid_loss: 0.0622067738420467762 epoch_valid_loss: 0.13844437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  epoch_valid_loss: 0.062706773842046762.	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32Ds.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.111458584666625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch_27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: %f 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.338844379884197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: %f 0.042814590554270476	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: Xf 0.004317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: Xf 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: Xf 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.05220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.05220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.1884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: Xf 0.042814590554270476, epoch_train_loss: 0.12242183658181374, epoch: 28	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1355783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.062705708887459649 epoch_valid_loss: 0.062705708887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: %f 0.062706773842047762 epoch_valid_loss: 0.062706773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: %f 0.042814590654270476 epoch_valid_loss: 0.042814590654270476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: \$\fiftin \text{ 0.0817793250083923} epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: \$\fiftin \text{ 0.06270570887459649} epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145853466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.16264. lr 2.3404e-04: 100%  test loss: \$\fiftin \text{ 0.06220773842047762}, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: \$\fiftin \text{ 0.0428184590554270476}, epoch_train_loss: 0.1224183658181374, epoch: 28 Saving at epoch_28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.176885905225845344 train_step: 13500, learning_rate: 0.0002120071803984038	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<11:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: Xf 0.04317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.13578379498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: Xf 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: Xf 0.05220773842047762 epoch_valid_loss: 0.05220773842047762 epoch_valid_loss: 0.05220773842047762 epoch_valid_loss: 0.05220773842047762 epoch_valid_loss: 0.042814590554276476 step_train_loss: 0.1884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: Xf 0.042814590554276476 epoch_valid_loss: 0.042814590554276476, epoch_train_loss: 0.1224183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1768859022584534 train_step: 13500, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lp 2.1338e-047: 1500%	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.135783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch_26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.062705708887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch_27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: %f 0.06220773842047762 epoch_valid_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_28 iter 477: train loss 0.18661. lr 2.2374e-04: 100%  test loss: %f 0.042814590554270476 epoch_valid_loss: 0.042814590554270476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch_28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.17768859902584534 train_step: 13500, learning_rate: 0.00022120971803984038 epoch_29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test_loss: %f 0.0350393446645251	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<11:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: \$\frac{\psi}{0.08437793250083923} = \text{poch_valid_loss: 0.04317793250083923}, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: \$\frac{\psi}{0.06270570887459649} = \text{poch_train_loss: 0.1356973075872934}, epoch: 26 step_train_loss: 0.1145853466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: \$\frac{\psi}{0.062209773842047762} = \text{poch_train_loss: 0.12931388097515167}, epoch: 27 step_train_loss: 0.05220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.05220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.042814599554270476, epoch_train_loss: 0.1242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.16788599025284534 train_step: 13500, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: \$\frac{\psi}{0.0620077368599025845344} train_step: 13500, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: \$\frac{\psi}{0.0620077368599025845344} train_step: 13500, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: \$\frac{\psi}{0.06200773685992586645251}, epoch_train_loss: 0.11676899780944053, epoch: 29	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<11:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]
step_train_loss: 0.1290904989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: \$\f 0.04317793250803923 epoch_valid_loss: 0.04317793250803923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.1357837949818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: \$\f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: \$\f 0.06220773842047762 epoch_valid_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: \$\f 0.042814590554270476 epoch_valid_loss: 0.042814590554270476, epoch_train_loss: 0.1224183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.1768859922584534 train_step: 13500, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: \$\f 0.0350393446645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]    94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: \$\fiftig{K}\$ 0.08437793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: \$\fiftig{K}\$ 0.06270570887459649 epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.15264. lr 2.3404e-04: 100%  test loss: \$\fiftig{K}\$ 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.062207738420447762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: \$\fiftig{K}\$ 0.06220773842047762, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_loead_16emb_32bs.pt step_train_loss: 0.177688590525846543 train_step: 13500, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: \$\fiftig{K}\$ 0.05280393446645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_loead_16emb_32bs.pt step_train_loss: 0.0350393446645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_loead_16emb_32bs.pt step_train_loss: 0.0559580317607412 train_step: 13000 learning_crate: 0.0002120373802253222532284	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:23<00:00, 5.82it/s]   478/478 [01:22<00:00, 5.82it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: Xf 0.084317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: Xf 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: Xf 0.06220773842047762 epoch_valid_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.052884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_valid_loss: 0.042814590554270476, epoch_train_loss: 0.1294183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt test loss: Xf 0.03503934466645251, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt test loss: Xf 0.03503934466645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1065956801760712 train_step: 13000, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: Xf 0.03503934466645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1065956801760712 train_step: 14000, learning_rate: 0.000211055837225392284 epoch 30 iter 477: train loss 0.12772. lr 2.2028e-04: 1000X	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]    94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_valld_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837949818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: %f 0.06270773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.1388443708417998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_valid_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: %f 0.042814590654270476 epoch_valid_loss: 0.042814590654270476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.17688590922584534 train_step: 13500, learning_rate: 0.00022120971803984038 epoch_29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: %f 0.0350393446645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch_29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1665956801766712 train_step: 14000, learning_rate: 0.0002120971803984038 epoch_29 iter 477: train loss 0.12222. lr 2.0928e-04: 100%  test loss: %f 0.031822561727695994	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:23<00:00, 5.82it/s]   478/478 [01:22<00:00, 5.82it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: Xf 0.084317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: Xf 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: Xf 0.06220773842047762 epoch_valid_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.052884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_valid_loss: 0.042814590554270476, epoch_train_loss: 0.1294183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt test loss: Xf 0.03503934466645251, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt test loss: Xf 0.03503934466645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1065956801760712 train_step: 13000, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: Xf 0.03503934466645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1065956801760712 train_step: 14000, learning_rate: 0.000211055837225392284 epoch 30 iter 477: train loss 0.12772. lr 2.2028e-04: 1000X	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:23<00:00, 5.82it/s]   478/478 [01:22<00:00, 5.82it/s]
step_train_loss: 0.1290904989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%  test loss: %f 0.04317793250083923 epoch_valld_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.135783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: %f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. lr 2.3404e-04: 100%  test loss: %f 0.06270773842047762 epoch_valid_loss: 0.062708773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.18461. lr 2.2374e-04: 100%  test loss: %f 0.042814590554270476 epoch_valid_loss: 0.042814590554270476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1768850922584534 train_step: 13500, learning_rate: 0.00022120971803984038 epoch 29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: %f 0.0350393446645251 epoch_valid_loss: 0.0350393446645251, epoch_train_loss: 0.116768997809404053, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.10659568011760712 train_step: 14000, learning_rate: 0.0002120971803984038 epoch 29 iter 477: train loss 0.12972. lr 2.0298e-04: 100%  test loss: %f 0.0328256172769994, epoch_train_loss: 0.1137090459171759, epoch: 30 Saving at epoch 30: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1512806415557866769944, epoch_train_loss: 0.103709459171759, epoch: 30 Saving	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:23<00:00, 5.82it/s]   478/478 [01:22<00:00, 5.82it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch_25 iter_477: train_loss 0.12187. lr 2.5433e-04: 100%  test_loss: #f .004317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving_at epoch_25: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1357837498188082 train_step: 12000, learning_rate: 0.0002532592516845316 epoch_26 iter_477: train_loss 0.14351. lr 2.4424e-04: 100%  test_loss: #f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466625214 train_step: 12500, learning_rate: 0.00024269198632935544 epoch_27 iter_477: train_loss 0.12624. lr 2.3404e-04: 100%  test_loss: #f 0.06220773842047762 epoch_valid_loss: 0.0627073842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.052884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_28 iter_477: train_loss 0.18461. lr 2.2374e-04: 100%  test_loss: #f 0.042814590554270476 epoch_valid_loss: 0.042814590554270476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving_at epoch_28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.176889902584374 train_step: 13500, learning_rate: 0.00022120971803984038 epoch_29 iter_477: train_loss_0.13608. lr 2.1338e-04: 100%  test_loss: #f 0.03503934466645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving_at epoch_29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt  test_loss: #f 0.031282561727695994 epoch_valid_loss: 0.10582561727695994, epoch_train_loss: 0.113709459171759, epoch: 30 Saving_at_epoch_29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt  test_loss: #f 0.031282561727695994 epoch_valid_loss: 0.031282561727695994, epoch_train_loss: 0.113709459171759, epoch: 30 Saving_at_epoch_29: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt  test_loss: #f 0.031282561727695994 ep	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [01:22<00:00, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]
step_train_loss: 0.12999994989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. Ir 2.5433e-04: 100% test loss: % ff 0.04317793250883933 epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_2head_l6embd_32bs.pt step_train_loss: 0.135783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. Ir 2.4424e-04: 100% test loss: % ff 0.0627076887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.11145858466652514 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train loss 0.12624. Ir 2.3404e-04: 100% test loss: % ff 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_valid_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch_valid_loss: 0.0428145990554270476 epoch_valid_loss: 0.0428145990554270476 epoch_valid_loss: 0.0428145990554270476 epoch_valid_loss: 0.0428145990554270476 epoch_valid_loss: 0.05360393446645251 epoch_valid_loss: 0.055609218454 train_step: 13500, learning_rate: 0.00022120971803984038 epoch_29 iter 477: train_loss_0.13608. Ir 2.1338e-04: 100% test_loss: % ff 0.05093303446645251 epoch_valid_loss: 0.0556956011760712 train_step: 13500, learning_rate: 0.00022120971803984038 epoch_29 iter_477: train_loss_0.13608. Ir 2.1338e-04: 100% test_loss: % ff 0.0509560911760712 train_step: 14000, learning_rate: 0.0002210971803984038 epoch_20 iter_477: train_loss_0.12608. Ir 2.1338e-04: 100% test_loss: % ff 0.0308303146645251, epoch_train_loss: 0.113769459171759, epoch: 30 Saving at epoch_20: ./cond_gpt/weights/Mone_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.016595609176699994, epoch_train_loss: 0.1137694591717599 epoch_3	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]
step_train_loss: 0.12999994989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. Ir 2.5433e-04: 100%  test loss: % ff 0.04317793250083923 epoch valid_loss: 0.04317793250083923 epoch valid_loss: 0.04317793250083923 epoch_valid_loss: 0.04317793250083923 epoch_valid_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch_25 iter 477: train loss 0.14351. Ir 2.4424e-04: 100%  test loss: % ff 0.06220778887459649 epoch_valid_loss: 0.06270570887459649 epoch_valid_loss: 0.06270570887459649 epoch_valid_loss: 0.06220773842047762 epoch_valid_loss: 0.06220773842047764 epoch_valid_loss: 0.0622077384204766 epoch_valid_loss: 0.0622077384204766 epoch_valid_loss: 0.002814590554270476 epoch_valid_loss: 0.0028145905542704766 epoch_valid_loss: 0.00281459054256177695994 epoch_valid_loss: 0.00281459055578615 train_step: 14000, learning_rate: 0.0002120971803984038 epoch_valid_loss: 0.0028175787695994 epoch_valid_loss: 0.004388365578768555 epoch_train_loss: 0.10855516615990066, epoch: 31	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<11:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   478/478 [01:20<00:00, 5.78it/s]
step_train_loss: 0.12999994989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477; train loss 0.12187. lr 2.5432e-04: 100%   test loss: % f0.043177937508833923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.1335783749818882 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477; train loss 0.14351. lr 2.442de-04: 100%   test loss: % f0.06276570887459649   epoch_valid_loss: 0.06276570887459649   epoch_valid_loss: 0.06276570887459649   epoch_valid_loss: 0.062708773842047762   epoch_valid_loss: 0.06220773842047762   epoch_valid_loss: 0.06220773842047762   epoch_valid_loss: 0.13884437084197998 train_step: 12000, learning_rate: 0.00024269198632935544   epoch_valid_loss: 0.06220773842047762   epoch_valid_loss: 0.06220773842047762   epoch_valid_loss: 0.06220773842047762   epoch_valid_loss: 0.06220773842047762   epoch_valid_loss: 0.08202773842047762   epoch_valid_loss: 0.08202773842047762   epoch_valid_loss: 0.08202773842047762   epoch_valid_loss: 0.08202773842047762   epoch_valid_loss: 0.06220773842047762   epoch_valid_loss: 0.08202773842047762   epoch_valid_loss: 0.08202773842047762   epoch_valid_loss: 0.082027278470476   epoch_valid_loss: 0.082027204762   epoch_valid_loss: 0.082027204769999   epoch_valid_loss: 0.0820272047699999   epoch_valid_loss: 0.082027204769999   epoch_valid_loss: 0.08202720476959999   epoch_valid_loss:	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [01:22<00:00, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:28<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.91it/s]   478/478 [01:22<00:00, 5.91it/s]   478/478 [01:20<00:52, 6.07it/s]   478/478 [01:20<00:00, 5.91it/s]
step_train_loss: 0.1299994989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5432e-04: 100%  test loss: % 0.043177937508033923 epoch valid_loss: 0.043177937508033923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: (room_gpt/weights/None_lengthsplit_50layer_zhead_16embd 32bs.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%  test loss: % 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.06270577842047762 epoch_valid_loss: 0.0627973842047762 epoch_valid_loss: 0.06220773842047762 epoch_valid_loss: 0.0482814590554270476 epoch_valid_loss: 0.0482814590554270476 epoch_valid_loss: 0.1588443708443708416. lr 2.23740e-04: 100%  test loss: % 0.06220773842047762 epoch_valid_loss: 0.05284534 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.136463. lr 2.23740e-04: 100%  test loss: % 0.05309393466452740476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1076885092284534 train_step: 13500, learning_rate: 0.00022120971803984038 epoch_29 iter 477: train loss 0.13608. lr 2.1338e-04: 100%  test loss: % 0.0330939346645251, epoch_train_loss: 0.11676899780944053, epoch: 29 Saving at epoch_92: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.1059568811769712 train_step: 14000, learning_rate: 0.0002120971803984038 epoch_90ch_valid_loss: 0.03128256172769994, epoch_train_loss: 0.1132709459171759, epoch: 30 Saving at epoch 30: ./cond_g	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<11:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   478/478 [01:20<00:00, 5.78it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train_loss 0.12187. lr 2.5433c-04: 100%[ test loss: %f 0.0317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: /cond_gpt/weights/None_lengthsplit_50layer_zhead_l6embd_32bs.pt step_train_loss: 0.133783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train_loss 0.1351. lr 2.4424c-04: 100%[ test loss: %f 0.06270570887459649, epoch_train_loss: 0.13569730758772934, epoch: 26 step_train_loss: 0.05270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.05270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.13588366652514 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 27 iter 477: train_loss 0.12624. lr 2.3404c-04: 100%[ test loss: %f 0.062270773842047762 epoch_valid_loss: 0.05220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13684437084157989 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train_loss 0.18461. lr 2.2374c-04: 100%[ test loss: %f 0.042814590554270476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.105859580925584534 train_step: 13500, learning_rate: 0.00022120971803984038 epoch_29 iter 477: train_loss 0.13668. lr 2.1338c-04: 100%[ test loss: %f 0.036303446665251, epoch_train_loss: 0.1167689978094063, epoch: 29 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.10659568011760712 train_step: 14000, learning_rate: 0.0002120971803984038 epoch_29 iter 477: train_loss 0.12792. lr 2.0298c-04: 100%[ test loss: %f 0.031828561727655994, epoch_train_loss: 0.1132709459171759, epoch: 30 Saving at epoch 29: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.1065956801760712 train	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [01:22<00:00, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:28<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.91it/s]   478/478 [01:22<00:00, 5.91it/s]   478/478 [01:20<00:52, 6.07it/s]   478/478 [01:20<00:00, 5.91it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%[ test loss: % 6.08317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: \cdotson dgtf\verlgibtf\/lose\lengthsplit_50layer_Zhead_ifeabd_32bs.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%[ test loss: % 6.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.1158883666652514 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 77 iter 477: train loss 0.12624. lr 2.3404e-04: 100%[ test loss: % 6.06220773842047762 epoch_valid_loss: 0.06220773842047762 epoch_valid_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.1616. lr 2.2374e-04: 100%[ test loss: % 6.042814590554270476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: \cdots \cdots \cdot \cdots \cdot \cd	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [01:22<00:00, 5.83it/s]   94/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   182/478 [00:26<00:52, 6.07it/s]   478/478 [01:20<00:00, 5.91it/s]   182/478 [00:31<00:51, 5.76it/s]   182/478 [00:31<00:51, 5.76it/s]   478/478 [01:22<00:00, 5.77it/s]
step_train_loss: 0.1299994989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. In 2.5433e-04: 100%  test loss: %f 0.04317793259083923 epoch_valid_loss: 0.04317793259083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: /coor_ggt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train_loss 0.14351. Tr_2.4424e-04: 100%  test loss: %f 0.06270570887459649 epoch_valid_loss: 0.06270570887459649 epoch_valid_loss: 0.06270570887459649 epoch_valid_loss: 0.06270570887459649 epoch_valid_loss: 0.108584366675214 train_step: 12500, learning_rate: 0.000242699198632935544 epoch_27 iter 477: train_loss 0.116242. In 2.3404e-04: 100%  test loss: %f 0.0622073842047762 epoch_valid_loss: 0.0622073842047762 epoch_valid_loss: 0.0622073842047762 epoch_valid_loss: 0.0622073842047762 epoch_valid_loss: 0.0622073842047762 epoch_valid_loss: 0.0622073842047762 epoch_valid_loss: 0.0622073842047676 epoch_valid_loss: 0.062884590554270476 epoch_valid_loss: 0.0628869022886542 train_step: 14000%  test loss: %f 0.0838846645251 epoch_valid_loss: 0.0848869022886542 train_step: 14000%  test loss: %f 0.084886452517 epoch_valid_loss: 0.084886645251, epoch_train_loss: 0.113270945917175 epoch_valid_loss: 0.0438812568768665 epoch_valid_loss: 0.0438812568768665 epoch_valid_loss: 0.04388125687686655 epoch_valid_loss: 0.0438812568768655 e	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   72/478 [00:12<00:00, 5.83it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   478/478 [01:20<00:00, 5.91it/s]   478/478 [01:22<00:00, 5.97it/s]   478/478 [01:22<00:00, 5.77it/s]   478/478 [01:22<00:00, 5.77it/s]
step_train_loss: 0.12909094989299774 train_step: 11500, learning_rate: 0.0002636687494788558 epoch 25 iter 477: train loss 0.12187. lr 2.5433e-04: 100%[ test loss: % 6.08317793250083923 epoch_valid_loss: 0.04317793250083923, epoch_train_loss: 0.14064275197466547, epoch: 25 Saving at epoch 25: \cdotson dgtf\verlgibtf\/lose\lengthsplit_50layer_Zhead_ifeabd_32bs.pt step_train_loss: 0.1335783749818802 train_step: 12000, learning_rate: 0.0002532592516845316 epoch 26 iter 477: train loss 0.14351. lr 2.4424e-04: 100%[ test loss: % 6.06270570887459649 epoch_valid_loss: 0.06270570887459649, epoch_train_loss: 0.1356973075872934, epoch: 26 step_train_loss: 0.1158883666652514 train_step: 12500, learning_rate: 0.00024269198632935544 epoch 77 iter 477: train loss 0.12624. lr 2.3404e-04: 100%[ test loss: % 6.06220773842047762 epoch_valid_loss: 0.06220773842047762 epoch_valid_loss: 0.06220773842047762, epoch_train_loss: 0.12931388097515167, epoch: 27 step_train_loss: 0.13884437084197998 train_step: 13000, learning_rate: 0.00023199823520207903 epoch 28 iter 477: train loss 0.1616. lr 2.2374e-04: 100%[ test loss: % 6.042814590554270476, epoch_train_loss: 0.12242183658181374, epoch: 28 Saving at epoch 28: \cdots \cdots \cdot \cdots \cdot \cd	478/478 [01:20<00:00, 5.95it/s]   50/478 [00:08<01:11, 5.99it/s]   478/478 [01:23<00:00, 5.72it/s]   72/478 [00:12<01:11, 5.65it/s]   478/478 [01:22<00:00, 5.83it/s]   94/478 [00:15<01:05, 5.91it/s]   478/478 [01:23<00:00, 5.72it/s]   116/478 [00:20<01:04, 5.64it/s]   478/478 [01:22<00:00, 5.82it/s]   138/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.78it/s]   478/478 [01:22<00:00, 5.91it/s]   182/478 [00:31<00:51, 5.76it/s]   182/478 [00:31<00:51, 5.76it/s]   478/478 [01:22<00:00, 5.77it/s]

test loss: %f 0.029598432248113333		
epoch_valid_loss: 0.029598432248113333, epoch_train_loss: 0.10103700118271876, epoch: 33		
Saving at epoch 33: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt		
step_train_loss: 0.08311779797077179 train_step: 16000, learning_rate: 0.00016696836159981425	226/478 [00:39<00:43,	
epoch 34 iter 477: train loss 0.10155. lr 1.6159e-04: 100%	478/478 [01:22<00:00,	5.78it/s]
test loss: %f 0.028342962644442363 epoch valid loss: 0.028342962644442363, epoch train loss: 0.09698281119184016, epoch: 34		
epoul_valu_ioss. v.22c3=22c2=4=4=2c0.; epoul_crail_ioss. corposecalifyid=4c1, epoul. 34 Saving at epoch 34: ./cond_gpt/weights/None lengthsplit 50layer_Blead 16embd 32bs.pt		
step train loss: 0.1051926240324974 train step: 16500, learning rate: 0.00015628542586217074	248/478 [00:43<00:40,	5.73it/s]
epoch 35 iter 477: train loss 0.06661. lr 1.5144e-04: 100%	478/478 [01:23<00:00,	
test loss: %f 0.04783162883379393		
epoch_valid_loss: 0.04783162883379393, epoch_train_loss: 0.09303848630121313, epoch: 35		
step_train_loss: 0.09395767003297806 train_step: 17000, learning_rate: 0.00014573198390292774	270/478 [00:46<00:35,	
epoch 36 iter 477: train loss 0.07351. lr 1.4141e-04: 100%	478/478 [01:22<00:00,	5.79it/s]
test loss: %f 0.0254517739473118		
epoch_valid_loss: 0.02354517739473118, epoch_train_loss: 0.09020787828200531, epoch: 36 Saving at epoch 36: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt		
saving at epoch so: ./com_gpt/weights/monm_lengthsplit_object_anead_noemmu_zcbs.pt step_train_loss: 0.086786827148456573 train_step: 17500_olawring_nate: 0.000133393180092872	292/478 [00:48<00:30,	6 11i+/s1
sep. 13 iter 477: train loss 0.18010. 1.3154e-04: 180%	478/478 [01:19<00:00,	
test loss: %f 0.03534271942313622	,, [,	
epoch_valid_loss: 0.03534271942313622, epoch_train_loss: 0.08673832455657765, epoch: 37		
step_train_loss: 0.08914818614721298 train_step: 18000, learning_rate: 0.0001251381531250835	314/478 [00:52<00:26,	6.09it/s]
epoch 38 iter 477: train loss 0.06940. lr 1.2186e-04: 100%	478/478 [01:19<00:00,	
test loss: %f 0.02087155810591799		
epoch_valid_loss: 0.02087155810591799, epoch_train_loss: 0.0823182707816113, epoch: 38		
Saving at epoch 38: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	1 22C/479 [00 FF-00 02	C 0011 ( 1
step_train_loss: 0.05500321090221405 train_step: 18500, learning_rate: 0.00011515874910140612 epoch 39 iter 477: train loss 0.06567. lr 1.1239e-04: 100%	336/478 [00:55<00:23,   478/478 [01:20<00:00,	
epocn 39 lter 4/7: train 1055 0.0050/. 1r 1.1239e-04: 100%  test loss: \$# 0.021883436544156738	476/476 [01:20000:00,	3.3/10/5]
epoch_valid_loss: 0.021083435544156738, epoch_train_loss: 0.08251159249660735, epoch: 39		
step_train_loss: 0.06773664802312851 train_step: 19000, learning_rate: 0.00010543070538687811	358/478 [01:01<00:20,	5.77it/s]
epoch 40 iter 477: train loss 0.09866. lr 1.0316e-04: 100%	478/478 [01:22<00:00,	
test loss: %f 0.018854224610280385		
epoch_valid_loss: 0.018854224610280385, epoch_train_loss: 0.07876986779163447, epoch: 40		
Saving at epoch 40: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt		
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05	380/478 [01:05<00:16,	5.78it/s]
epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%	478/478 [01:22<00:00,	5.801t/s]
test loss: %F 0.01730320260308131		
step_train_loss: 0.0539277166788578 train_step: 19500, learning_rate: 9.598276245103928e-05	380/478 [01:05<00:16,	
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%	380/478 [01:05<00:16,   478/478 [01:22<00:00,	
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100% test loss: %f 0.01730320260308131		
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train_loss 0.06956. lr 9.4187e-05: 100% test_loss: %f 0.01730320260308131 epoch_valid_loss: 0.017303202630308131, epoch_train_loss: 0.07846453028368651, epoch: 41		
step_train_loss: 0.0539277166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100% test loss: %f 0.01730320260308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_3Zbs.pt	478/478 [01:22<00:00,	5.80it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: \$f 0.01730320260308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_15embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05	478/478 [01:22<00:00,	5.80it/s] 5.65it/s]
step_train_loss: 0.0539277166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100% test loss: %f 0.01730320260308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_3Zbs.pt	478/478 [01:22<00:00,	5.80it/s] 5.65it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100% test loss: %f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0833921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100% test loss: %f 0.018228005781700765	478/478 [01:22<00:00,	5.80it/s] 5.65it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: %f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%   test loss: %f 0.018228805781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06063132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05	478/478 [01:22<00:00]   402/478 [01:10<00:13]   478/478 [01:23<00:00]   424/478 [01:12<00:08]	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%  test loss: %f 0.04730320260308131, epoch_train_loss: 0.07846453028358651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08332921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%  test loss: %f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06063132360577588 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch 43 iter 477: train loss 0.07765. lr 7.7128e-05: 100%  epoch_dail_train_loss 0.0606755. lr 7.7128e-05: 100%  epoch_dail_train_loss 0.06067813260577588 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_dail_train_loss 0.07765. lr 7.7128e-05: 100%	478/478 [01:22<00:00, 402/478 [01:10<00:13, 478/478 [01:23<00:00,	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: %f 0.01730320260308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08329291713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%   test loss: %f 0.018228005781700765, epoch train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06063132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.018228005781700765. lr 7.7128e-05: 100%   test loss: %f 0.01730475913829807	478/478 [01:22<00:00]   402/478 [01:10<00:13]   478/478 [01:23<00:00]   424/478 [01:12<00:08]	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: %f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5593e-05: 100%   test loss: %f 0.018228805781700765 epoch_valid_loss: 0.018228805781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06605132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_43 iter 477: train loss 0.07765. lr 7.7128e-05: 100%   test loss: %f 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43	478/478 [01:22<00:00]   402/478 [01:10<00:13]   478/478 [01:23<00:00]   424/478 [01:12<00:08]	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: \$f 0.017303202669308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832927173590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.066405. lr 8.5503e-05: 100%   test loss: \$f 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.0862313260577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch 43 iter 477: train loss 0.07765. lr 7.7128e-05: 100%   test loss: \$f 0.017130475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	478/478 [01:22<00:00]   402/478 [01:10<00:13]   478/478 [01:23<00:00]   424/478 [01:12<00:08]   478/478 [01:22<00:00]	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100% test loss: %f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100% test loss: %f 0.018228005781700765 epoch_valid_loss: 0.0182280805781700765, epoch train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06063132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.00774172767996788 train_step: 210000, learning_rate: 6.959503319984364e-05	478/478 [01:22<00:00]   402/478 [01:10<00:13]   478/478 [01:23<00:00]   424/478 [01:12<00:08]   478/478 [01:22<00:00]   446/478 [01:16<00:05]	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%  test loss: %f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5593e-05: 100%  test loss: %f 0.018228805781700765 epoch_valid_loss: 0.018228805781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06663132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_43 iter 477: train loss 0.07765. lr 7.7128e-05: 100%  test loss: %f 0.017130475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.077472767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch 44 iter 477: train loss 0.04170. lr 6.5086e-05: 100%	478/478 [01:22<00:00]   402/478 [01:10<00:13]   478/478 [01:23<00:00]   424/478 [01:12<00:08]   478/478 [01:22<00:00]	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: Xf 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08329297173590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%   test loss: Xf 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.0663132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch 44 iter 477: train loss 0.04170. lr 6.9066e-05: 100%   test loss: Xf 0.01637202553131567	478/478 [01:22<00:00]   402/478 [01:10<00:13]   478/478 [01:23<00:00]   424/478 [01:12<00:08]   478/478 [01:22<00:00]   446/478 [01:16<00:05]	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100% test loss: \$f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.083239291713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.086205. lr 8.5503e-05: 100% test loss: \$f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.08663132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_d1 iter 477: train loss 0.07765. lr 7.7128e-05: 100% test loss: \$f 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_44 iter 477: train loss 0.04470. lr 6.9086e-05: 100% test loss: \$f 0.016327320253131567 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.070916213599583, epoch: 44 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	478/478 [01:22<00:00]   402/478 [01:10<00:13]   478/478 [01:23<00:00]   424/478 [01:12<00:08]   478/478 [01:22<00:00]   446/478 [01:16<00:05]	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100% test loss: %f 0.01730320260308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08329291713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100% test loss: %f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.0663132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.070473344733400154114 train_step: 21500, learning_rate: 6.153799489525387e-05	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:08],   478/478 [01:22<00:00],   446/478 [01:16<00:05],   478/478 [01:22<00:00],   468/478 [01:21<00:01],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%  test loss: %f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%  test loss: %f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.0663132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch 43 iter 477: train loss 0.07765. lr 7.7128e-05: 100%  test loss: %f 0.01730475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.0747127767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch 44 iter 477: train loss 0.04170. lr 6.9086e-05: 100%  test loss: %f 0.016327320253131567 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt  test loss: %f 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt  step_train_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt  step_train_loss: 0.01437433400154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch 45 iter 477: train loss 0.03274. lr 6.1399e-05: 100%	478/478 [01:122<00:00],   492/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:08],   478/478 [01:22<00:00],   446/478 [01:16<00:05],   478/478 [01:22<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100% test loss: \$f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08329291713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06605. lr 8.5503e-05: 100% test loss: \$f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003288075607, epoch: 42 step_train_loss: 0.08663132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch 43 iter 477: train loss 0.07765. lr 7.7128e-05: 100% test loss: \$f 0.017130475913829827, epoch_train_loss: 0.07145554593761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch 44 iter 477: train_loss 0.04170. lr 6.9086e-05: 100% test loss: \$f 0.016327320253131567, epoch_train_loss: 0.070401276996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.070401276996788 train_step: 21500, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.070401276996788 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch_valid_loss: 0.08147433400154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch_valid_loss: 0.0	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:08],   478/478 [01:22<00:00],   446/478 [01:16<00:05],   478/478 [01:22<00:00],   468/478 [01:21<00:01],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%  test loss: %f 0.01730320260308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08329291713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%  test loss: %f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06063132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_43 iter 477: train loss 0.07765. lr 7.7128e-05: 100%  test loss: %f 0.01730475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch_43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_44 iter_477: train_loss 0.04170. lr 6.9086e-05: 100%  test loss: %f 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch_44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08147434340815414 train_step: 215000, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch_44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08147433408154314 train_step: 215000, learning_rate: 6.553799489525387e-05 epoch_valid_loss: 0.0159001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:08],   478/478 [01:22<00:00],   446/478 [01:16<00:05],   478/478 [01:22<00:00],   468/478 [01:21<00:01],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: \$f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08329291713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.086205. lr 8.5503e-05: 100%   test loss: \$f 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.0862313260577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch 43 iter 477: train loss 0.07765. lr 7.7128e-05: 100%   test loss: \$f 0.01730475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 42 step_train_loss: 0.0077130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_43 iter 477: train loss 0.044700. lr 6.9086e-05: 100%   test loss: \$f 0.016327320253131567 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt step_train_loss: 0.084147433408154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch_45 iter 477: train loss 0.03741. lr 6.1399e-05: 100%   test loss: \$f 0.0159001647552941 epoch_valid_loss: 0.0159001647552941, epoch_train_loss: 0.06880093894394023, epoch: 45 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_zhead_16embd_32bs.pt	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:00],   424/478 [01:22<00:00],   446/478 [01:22<00:00],   446/478 [01:22<00:00],   468/478 [01:21<00:01],   478/478 [01:23<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s] 5.84it/s] 5.73it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: Xf 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08329297173590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%   test loss: Xf 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.0862313260577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554593761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.070916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08147433400154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.0709106213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08147433400154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch_valid_loss: 0.0159001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt epoch_valid_loss: 0.0159001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:08],   478/478 [01:22<00:00],   446/478 [01:16<00:05],   478/478 [01:22<00:00],   468/478 [01:21<00:01],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s] 5.84it/s] 5.73it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%  test loss: %f 0.01730320260308131 epoch_valid_loss: 0.04730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832929171590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%  test loss: %f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06063132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.08147433400154314 train_step: 11500, learning_rate: 6.153799489525387e-05 epoch_valid_loss: 0.0150901647552941 epoch_valid_loss: 0.0150901647552941 epoch_valid_loss: 0.0150901647552941 epoch_valid_loss: 0.0150901647552941 epoch_valid_loss: 0.0150901647552961 epoch_valid_loss: 0.0150901647552961 epoch_valid_loss: 0.0150901647552961 epoch_valid_loss: 0.0150901647552961 epoch_valid_loss: 0.0150901647552961 epoch_valid_loss: 0.0150901647552961 epoch_valid_loss: 0.0150901647	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:00],   424/478 [01:22<00:00],   446/478 [01:22<00:00],   446/478 [01:22<00:00],   468/478 [01:21<00:01],   478/478 [01:23<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s] 5.84it/s] 5.73it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%   test loss: \$f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832927173590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.0862052 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.086208781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003288075607, epoch: 42 step_train_loss: 0.0862313360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567 epoch_valid_loss: 0.016327320253131567 epoch_valid_loss: 0.016327320253131567 epoch_valid_loss: 0.08147433400154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch_45 iter 477: train_loss 0.03274. lr 6.1399e-05: 100%  test_loss: \$f 0.0150901647552941 epoch_valid_loss: 0.0159001647552941 epoch_valid_loss: 0.0159001647552941 epoch_valid_loss: 0.0159001647552941 epoch_valid_loss: 0.0159001647552941 epoch_valid_loss: 0.016665317162967944 epoch_valid_loss: 0.016665317162967944 epoch_valid_loss: 0.016665317162967944, epoch_train_loss: 0.06880697374516, epoch: 46 epoch_valid_loss: 0.016665317162967944, epoch_train_loss: 0.06880897374516, epoch: 46 epoch_valid_loss: 0.016665317162967944, epoch_train_loss: 0.06880897374516, epoch: 46	478/478 [01:22<00:00],   402/478 [01:10<00:13,   478/478 [01:23<00:00],   424/478 [01:23<00:00],   478/478 [01:22<00:00],   446/478 [01:16<00:05,   478/478 [01:22<00:00],   468/478 [01:23<00:00],   478/478 [01:23<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.84it/s] 5.73it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%  test loss: %f 0.01730320260308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%  test loss: %f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06065132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.059001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt epoch_valid_loss: 0.0159001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt test loss: %f 0.016065317162967944, epoch_train_loss: 0.0688089767374516, epoch: 46 step_train_loss: 0.066577377	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:00],   424/478 [01:22<00:00],   446/478 [01:22<00:00],   446/478 [01:22<00:00],   468/478 [01:21<00:01],   478/478 [01:23<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.84it/s] 5.73it/s] 5.42it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. Ir 9.4187e-05: 100%  test loss: %f 0.017303202669308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_2head_l6embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. Ir 8.5503e-05: 100%  test loss: %f 0.018228005781700765 epoch_valid_loss: 0.018228005781700765 epoch_valid_loss: 0.018228005781700765 epoch_valid_loss: 0.018228005781700765 epoch_valid_loss: 0.017130475913829827 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 42 step_train_loss: 0.077765. Ir 7.7128e-05: 100%  test loss: %f 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch_43: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt step_train_loss: 0.0747172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_44 iter_477: train_loss 0.04170. Ir 6.9086e-05: 100%  test loss: %f 0.016327320253131567 epoch_train_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch_44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt test_loss: %f 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch_44: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt test_loss: %f 0.01509001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch_45: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt test_loss: %f 0.01509001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch_45: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.pt test_loss: %f 0.016065317162967044, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch_45: ./cond_gpt/weights/None_lengthsplit_50layer_2head_16embd_32bs.	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:08],   478/478 [01:22<00:00],   446/478 [01:16<00:05],   478/478 [01:22<00:00],   468/478 [01:21<00:01],   478/478 [01:23<00:00],   478/478 [01:28<00:00],   478/478 [01:28<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.84it/s] 5.73it/s] 5.42it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.66956. lr 9.4187e-05: 100% test loss: % 0.0173022063088131, epoch_train_loss: 0.07846453028368651, epoch: 41 Saving at epoch 41: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.0837927171590627 train_step: 20000, learning_rate: 8.68429465931604e-05 epoch 42 iter 477: train loss 0.66405. lr 8.5503e-05: 100% test loss: % 0.01822800578170675. poch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.085228005781706755, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06663132360577533 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch 43 iter 477: train loss 0.07755. lr 7.7128e-05: 100% test loss: % 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch 43 iter 477: train loss 0.047107. lr 6.09606e-05: 100% test loss: % 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt test loss: % 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt test loss: % 0.016327320253131567, epoch_train_loss: 0.068880693894394023, epoch: 45 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt test loss: % 0.0159001647552941 epoch_train_loss: 0.068880693894394023, epoch: 45 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt epoch_d6 iter 477: train_loss 0.05859. lr 5.4087e-05: 100% test loss: % 0.01506757105797763030 epoch_d7 iter 477: train_loss 0.058532. lr 4.7170e-05: 100% test loss: % 0.01506767165977763333, epoch_train_loss: 0.06667322512700467, epoch: 47	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:08],   478/478 [01:22<00:00],   446/478 [01:16<00:05],   478/478 [01:22<00:00],   468/478 [01:21<00:01],   478/478 [01:23<00:00],   478/478 [01:28<00:00],   478/478 [01:28<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.84it/s] 5.73it/s] 5.42it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. lr 9.4187e-05: 100%  test loss: % 6.08730260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 saving at epoch 41: /cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5593e-05: 100%  test loss: % 6.081228007581700765 epoch_valid_loss: 0.081228007581700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.0606513276007583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.01130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 42 step_train_loss: 0.0606313260075838 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.017130475913829927 epoch_valid_loss: 0.017130475913829927, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.07173047275995788 train_step: 210000, learning_rate: 6.959503319984364e-05 epoch_valid_loss: 0.016327320253131567 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.016327320253131567 epoch_valid_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt epoch_valid_loss: 0.01606531716296752041 epoch_valid_loss: 0.01606531716296752041, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt epoch_doss: 0.016065317162967304 epoch_valid_loss: 0.016065317162967304 epoch_valid_loss: 0.016065317162967344 epoch_valid_loss: 0.016065317162967304 epoch_valid_loss: 0.016065317162967304 epoch_valid_loss: 0.016065317162967304 epoch_	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:23<00:00],   478/478 [01:22<00:00],   446/478 [01:22<00:00],   478/478 [01:22<00:00],   468/478 [01:22<00:00],   478/478 [01:23<00:00],   478/478 [01:28<00:00],   478/478 [01:22<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s] 5.84it/s] 5.73it/s] 5.77it/s]
step_train_loss: 0.053972166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06056. lr 9.4187e-05: 100%  test loss: %f 0.01730320260308131, epoch_train_loss: 0.07846453028368651, epoch: 41 saving at epoch 41: /cond_gpt/weights/Mone_lengthsplit_50130yer_2/head_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. lr 8.5503e-05: 100%  test loss: %f 0.018228005781700765, epoch_train_loss: 0.0759603208075607, epoch: 42 step_train_loss: 0.083223095781700765, epoch_train_loss: 0.07596032080075607, epoch: 42 step_train_loss: 0.018228005781700765, epoch_train_loss: 0.07596032080075607, epoch: 42 step_train_loss: 0.018228005781300765, epoch_train_loss: 0.07596032080075607, epoch: 42 step_train_loss: 0.0187230007831700765, epoch_train_loss: 0.07596032080075607, epoch: 43 step_train_loss: 0.0187230007581700765, epoch_train_loss: 0.07145554503761076, epoch: 43 saving at epoch 43: /cond_gpt/weights/None_lengthsplit_50layer_2/head_16embd_32/bs.pt step_train_loss: 0.074172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch 44: iter 477: train loss 0.04170. lr 6.9086e-05: 100%  test loss: %f 0.016327320253131567, epoch_train_loss: 0.0703016213599583, epoch: 44 saving at epoch 43: /cond_gpt/weights/None_lengthsplit_50layer_2/head_16embd_32bs.pt step_train_loss: 0.08147433400154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch 45: iter 477: train loss 0.03272. lr 6.1399e-05: 100%  test loss: %f 0.0159001647552041 epoch_valid_loss: 0.0159001647552041, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch 43: /cond_gpt/weights/None_lengthsplit_50layer_2/head_16embd_32bs.pt epoch 45: iter 477: train loss 0.003833. lr 4.7170e-05: 100%  test loss: %f 0.0159001647552041, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch 43: /cond_gpt/weights/None_lengthsplit_50layer_2/head_16embd_32bs.pt test loss: %f 0.015067671697276333 epoch_valid_loss	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:23<00:00],   424/478 [01:22<00:00],   446/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:23<00:00],   478/478 [01:23<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.84it/s] 5.73it/s] 5.77it/s] 5.78it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.66956. Ir 9.4187e-05: 100%  test loss: % 0.01730320260308131 epoch_valid_loss: 0.01730320260308131, epoch_train_loss: 0.078464530283368551, epoch: 41 saving at epoch 41: /cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. Ir 8.5503e-05: 100%  test loss: % 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.082328005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.018238005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.018238005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.017130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.01517130475913829827, epoch_train_loss: 0.07145554503761076, epoch: 43 Saving at epoch 43: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 44 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.016327320253131567, epoch_train_loss: 0.07030916213599583, epoch: 45 Saving at epoch 44: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.016065317162067944, epoch_train_loss: 0.06888093894394023, epoch: 45 saving at epoch 45: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt test loss: %f 0.015067671697776333, epoch_train_loss: 0.06667322512700467, epoch: 47 Saving at epoch 47: ./cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt test loss: &f 0.015067671697776333, epoch_train_loss: 0.06667322512700467, epoch: 47 Saving at epoch 47: ./cond_	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:23<00:00],   478/478 [01:22<00:00],   446/478 [01:22<00:00],   478/478 [01:22<00:00],   468/478 [01:22<00:00],   478/478 [01:23<00:00],   478/478 [01:28<00:00],   478/478 [01:22<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.84it/s] 5.73it/s] 5.77it/s] 5.78it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. Ir 9.4187e-05: 100X  test loss: Xf 0.0179320260308131 epoch valid_loss: 0.01730320260308131, epoch_train_loss: 0.0784645302838368651, epoch: 41 Saving at epoch 41: /road gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32Ds.pt step_train_loss: 0.0832921713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.06405. Ir 8.5503e-05: 100X  test_loss: Xf 0.018228005781700765 epoch_valid_loss: 0.018228005781700765, epoch_train_loss: 0.07596003208075607, epoch: 42 step_train_loss: 0.06631332560577633 train_step: 20000, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.01730475913829827, epoch_train_loss: 0.077455554503761076, epoch: 43 step_train_loss: 0.006013730203313567 epoch_valid_loss: 0.017190475913829827, epoch_train_loss: 0.077455554503761076, epoch: 43 Saving at epoch_43: /road_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32Ds.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_44 iter_477: train_loss 0.04170. Ir 6.9086e-05: 100X  test_loss: Xf 0.016327320253131567, epoch_train_loss: 0.0709391621359583, epoch: 44 Saving at epoch_44: /road_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32Ds.pt step_train_loss: 0.08147433400154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch_45: /road_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32Ds.pt step_train_loss: 0.0816747552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving at epoch_45: /road_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32Ds.pt test_loss: Xf 0.0150001647555941, epoch_train_loss: 0.06880893894394023, epoch: 45 Saving at epoch_45: /road_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32Ds.pt test_loss: Xf 0.0150001647559401, epoch_train_loss: 0.06880893894394023, epoch: 45 Saving at epoch_45: /road_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32Ds.pt step_train_loss: 0.0806767169	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:23<00:00],   424/478 [01:22<00:00],   446/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:23<00:00],   478/478 [01:23<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.84it/s] 5.73it/s] 5.77it/s] 5.78it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. 1r 9.4187e-05: 100%  stept loss: 47 0.0173837063086131 epoch_valid_loss: 0.01738320263086131, epoch_train_loss: 0.07846453028368651, epoch: 41 saving at epoch 41: /cond_ggt/weights/Mone_lengthsplit_50layer_zhead_16embd_3zbs.pt step_train_loss: 0.0832971713590622 train_step: 20000, learning_rate: 8.684294645931604e-05 epoch 42 iter 477: train loss 0.064051. rs.5503e-05: 100%  test loss: 48 0.018228005781700765 epoch 43 loss: 0.018228005781700765, epoch_train_loss: 0.07556903208075607, epoch: 42 step_train_loss: 0.06063132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch 43 iter 477: train loss 0.07765. 1r 7.7128e-05: 100%  test loss: 48 0.0171307597918279827 epoch_valid_loss: 0.017310757918329827, epoch_train_loss: 0.07145554593761076, epoch: 43 Saving at epoch 43: /cond_ggt/weights/Mone_lengthsplit_50layer_zhead_16embd_3zbs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959593319984364e-05 epoch 43 iter 477: train loss 0.04170. lr 6.9086e-05: 100%  test loss: 48 0.016237230253131567 epoch_valid_loss: 0.016237230253131567 epoch_valid_loss: 0.01827320253131567 epoch_valid_loss: 0.01827320253131567 epoch_valid_loss: 0.01827320253131567 epoch_valid_loss: 0.01827320253131567 epoch_valid_loss: 0.01827320253131567 epoch_valid_loss: 0.018040147552941, epoch_train_loss: 0.06880693894394023, epoch: 44 Saving at epoch_44: /cond_ggt/weights/Mone_lengthsplit_50layer_zhead_16embd_3zbs.pt epoch_valid_loss: 0.018066371762967944 epoch_valid_loss: 0.018066371762967944 epoch_valid_loss: 0.0180663717629767944 epoch_valid_loss: 0.0180663717629767944 epoch_valid_loss: 0.08093808096079601 train_step: 22000, learning_rate: 3.38934603823668e-05 epoch_45 iter 477: train_loss 0.068077377676923866 train_step: 22000, learning_rate: 3.38934603823668e-05 epoch_45 iter 477: train_loss 0.0680771767444 epoch_valid_loss: 0.0180667371629767344 epoch_valid_loss: 0.01806	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:12<00:00],   424/478 [01:22<00:00],   446/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:23<00:00],   478/478 [01:23<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.84it/s] 5.73it/s] 5.77it/s] 5.88it/s] 5.78it/s]
step_train_loss: 0.0539272166788578 train_step: 19500, learning_rate: 9.598276245103928e-05 epoch 41 iter 477: train loss 0.06956. Ir 9.4187e-05: 100X  test loss: Xf 0.0179320260308131 epoch_valid_loss: 0.0179320260308131 epoch_valid_loss: 0.0179320260308131, epoch_train_loss: 0.07846453028386651, epoch: 41 Saving at epoch 41: /cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.0832921713590622 train_step: 20080, learning_rate: 8.684294645931604e-05 epoch_42 iter 477: train_loss 0.06405. Ir 8.5593e-05: 100X  test_loss: Xf 0.018220065781700765 epoch_valid_loss: 0.018220065781700765 epoch_valid_loss: 0.018220065781700765, epoch_train_loss: 0.075960032008075607, epoch: 42 step_train_loss: 0.0663132360577583 train_step: 20500, learning_rate: 7.803836597700791e-05 epoch_valid_loss: 0.017130475913829827, epoch_train_loss: 0.077457554503761076, epoch: 43 Saving at epoch_43: /cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.0774172767996788 train_step: 21000, learning_rate: 6.959503319984364e-05 epoch_44 iter_477: train_loss 0.04170. Ir 6.79086e-05: 100X  test_loss: Xf 0.016327320253131567, epoch_train_loss: 0.0709301621359583, epoch: 44 Saving_at_epoch_44: /cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt step_train_loss: 0.08147433400154114 train_step: 21500, learning_rate: 6.153799489525387e-05 epoch_45 iter_477: train_loss_0.08274. Ir 6.1399e-05: 100X  test_loss: Xf 0.0159001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving_at_epoch_45: /cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt epoch_45 iter_477: train_loss_0.05859. Ir 6.74807e-05: 100X  test_loss: Xf 0.0159001647552941, epoch_train_loss: 0.06880693894394023, epoch: 45 Saving_at_epoch_45: /cond_gpt/weights/None_lengthsplit_50layer_Zhead_16embd_32bs.pt epoch_45 iter_477: train_loss_0.05859. Ir 6.74807e-05: 100X  test_loss: Xf 0.0159001647552941, epoch_train_loss: 0.06880693894394023, epoch: 46 Saving_at_epoch_45: ./cond_gpt/weights/None_lengthsplit_50	478/478 [01:22<00:00],   402/478 [01:10<00:13],   478/478 [01:23<00:00],   424/478 [01:23<00:00],   424/478 [01:22<00:00],   446/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:23<00:00],   478/478 [01:23<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],   478/478 [01:22<00:00],	5.80it/s] 5.65it/s] 5.70it/s] 6.26it/s] 5.82it/s] 5.42it/s] 5.79it/s] 5.84it/s] 5.73it/s] 5.77it/s] 5.80it/s] 5.80it/s] 5.80it/s]

		26807 train_step: 23000		56/478 [00:09<01:12		
epoch 49 iter 4//: t test loss: %f 0.0171		0.04646. lr 4.0000e-05:	100%	478/478 [01:22<00:06	, 5./81t/s]	
			ss: 0.06294228218760825, epoch: 49			
				78/478 [00:13<01:09	5 79i±/s1	
	step_train_loss: 0.04153740033507347 train_step: 23500, learning_rate: 4e-05 epoch 50 iter 477: train loss 0.05615. lr 4.0000e-05: 100%					
test loss: %f 0.0150				478/478 [01:23<00:06	,,	
			ss: 0.06510838146114947, epoch: 50			
		89499 train_step: 24000		100/478 [00:17<01:09	, 5.76it/s]	
		.14123. lr 4.0000e-05:	00%	478/478 [01:22<00:06	, 5.78it/s]	
test loss: %f 0.0131						
			ss: 0.063196132692146, epoch: 51			
			_50layer_2head_16embd_32bs.pt			
		848526 train_step: 2450		122/478 [00:20<00:59		
		0.03850. lr 4.0000e-05:	.00%	478/478 [01:19<00:06	, 6.05it/s]	
test loss: %f 0.0180			0 0C347FF44049444CC			
epocn_valid_loss: 0.	01000942//0 6107360134E	88013 train_step: 25000	ss: 0.06317551404814466, epoch: 52	144/478 [00:23<00:5	6 06i+/e1	
		0.06003. lr 4.0000e-05:		478/478 [01:19<00:00		
test loss: %f 0.0137			00%	470/470 [01:15/00:00	, 0.0310/3]	
			ss: 0.061785944523172896, epoch: 53			
		58818 train_step: 25500		166/478 [00:28<00:51	. 6.10it/sl	
		.08961. lr 4.0000e-05:		478/478 [01:19<00:00		
test loss: %f 0.0150	97079953799	1645				
epoch_valid_loss: 0.	01509707995	3799645, epoch_train_lo	ss: 0.06042488103257063, epoch: 54			
step_train_loss: 0.1	21590599417	68646 train_step: 26000	learning_rate: 4e-05	188/478 [00:30<00:47		
		0.17466. lr 4.0000e-05:	.00%	478/478 [01:18<00:06	, 6.09it/s]	
test loss: %f 0.0145						
			ss: 0.06130679169178757, epoch: 55			
step_train_loss: 0.0	55161878466	60614 train_step: 26500	learning_rate: 4e-05	210/478 [00:34<00:44		
epoch 56 iter 477: t	rain loss 0	0.03525. lr 4.0000e-05:	100%	478/478 [01:18<00:06	, 6.05it/s]	
test loss: %f 0.0129	03952/382/0 01200205272	99370464	0 0C0993CAFAA00309F			
Enving at anoth F6.	/cond_apt/	weights (None lengther)	ss: 0.06088264544093085, epoch: 56 : 50layer 2head 16embd 32bs.pt			
		.86462 train_step: 27006		232/478 [00:38<00:46	6 13i+/cl	
		.05240. lr 4.0000e-05:		478/478 [01:18<00:00		
test loss: %f 0.0146	89087841866	745	COON	770/770 [01:10:00:00	, 0.0/10/5]	
			ss: 0.0604975133312079, epoch: 57			
			t_50layer_2head_16embd_32bs.pt			
		186462 train_step: 2700		232/478 [00:38<00:4		
test loss: %f 0.0146		0.05240. lr 4.0000e-05:	100%	478/478 [01:18<00:0	0, 6.0/1T/S]	
			ss: 0.0604975133312079, epoch: 57			
sten train loss: 0.	4281070455	908775 train sten: 2750	learning rate: 4e-05	254/478 [00:41<00:3	6 6 07it/s]	
step_train_loss: 0.04281070455908775 train_step: 27500, learning_rate: 4e-05 epoch 58 iter 477: train loss 0.08268. lr 4.0000e-05: 100%					0, 6.08it/s]	
test loss: %f 0.0138				1, [2212210010	.,,.,	
			ss: 0.059904210863705086, epoch: 58			
	step_train_loss: 0.037258751690387726 train_step: 28000, learning_rate: 4e-05					
epoch 59 iter 477: train loss 0.05775. lr 4.0000e-05: 100%					0, 6.05it/s]	
test loss: %f 0.0129						
			ss: 0.058687997685753154, epoch: 59			
			t_50layer_2head_16embd_32bs.pt			
step_train_loss: 0.04941820725798607 train_step: 28500, learning_rate: 4e-05 epoch 60 iter 477: train loss 0.08247. lr 4.0000e-05: 100%					9, 6.01it/s]	
epoch 60 iter 477: t test loss: %f 0.0137			100%	478/478 [01:19<00:0	0, 5.99it/s]	
			ss: 0.05798212577156317, epoch: 60			
epocii_valld_loss: 0.	013/0403685	//Jo/6//, epocn_crain_i	33. 0.03/302123//13031/, epoch: 00			
Results Table:						
n_layer n_head	n_embd	validation_loss	time_per_epoch			
50 2	16	1.3816234270731609	97.56907749176025			
50 2	16	1.0268689802399389	86.88305497169495			
50 2	16		86.88909959793091			
50 2	16		87.81722927093506			
50 2	16		87.01436400413513			
50 2	16		87.13222980499268			
50 2	16		87.5953061580658			
50 2	16	0.4480134127316652	86.93578743934631			
50 2	16		87.4491195678711			
50 2	16	0.30689116374210074				
50 2	16		87.27850127220154			
50 2 50 2	16 16	0.20577994264938212 0.16961453151371744				
50 2	16 16	0.169614531513/1/44 0.13510421346183177				
50 2	16 16	0.13510421346183177				
50 2	16	0.1352660658734816				
50 2	16	0.08133871956831878				

```
n_embd
16
16
16
                                                                                                                             validation loss
                                                                                                                             1.3816234270731609
1.0268689802399389
0.8456329692293096
                                                                                                                                                                                                         97.56907749176025
86.88305497169495
86.88909959793091
   0.7738873450844376
                                                                                                                                                                                                          87.81722927093506
                                                                                     16
16
16
16
16
16
                                                                                                                                                                                                         87.01436400413513
87.13222980499268
87.5953061580658
                                                                                                                             0.7046002171657704
0.6336997482511733
                                                                                                                             0.5708381583293279

      0.4480134127316652
      86.93578743934631

      0.38839665331222395
      87.4491195678711

      0.30689116374210074
      87.3046486377716

      0.24817123264074326
      87.27850127220154

      0.20577994264938212
      87.33934092521667

      0.16961453151371744
      87.21852445602417

      0.13510421346183177
      87.15836668014526

                                                                                     0.11595657950750103 87.0739336013794
0.1352660658734816 86.48729348182678
0.08133871956831878 86.01002764701843
0.11203097965982226 87.3331651687622
                                                                                                                             0.06979568711834925 87.21651434898376
0.06352007120019859 87.53285455703735
0.12730575766828325 86.09204387664795
                                                                                                                           0.12736575766628325 86.09264387664795 
0.6633081538246918 86.77224493026733 
0.05981088084755121 83.86673831939697 
0.047140366604758635 83.62232065200806 
0.04317793250083923 83.46513751174927 
0.06270570887459649 86.61711764335632 
0.06220773842047762 85.15019965171814 
0.042814590554270476 86.72601270675659 
0.0350393446645251 85.14984083175659 
0.03103961273768004 85.21649136007

    0.0356393446645251
    85.14984083175659

    0.031282761727695994
    85.82515001269997

    0.03438812578786855
    84.0733597278595

    0.03687326258255376
    85.95911145210266

    0.0255984322443113333
    85.35100626945496

    0.0263842962644442363
    85.82388114929199

    0.04783162883379939
    86.15980648994446

    0.02354517739473118
    85.59817576408386

                                                                                                                            0.05981088084755121 83.86673831939697
  0.047140366604758635 83.62232065200806
0.04317793250083923 83.40511751174927
0.06270570887459649 86.61711764335632
0.06220773842047762 85.15019965171814
                                                                                                                           0.042814590554270476 86.72601270675659
0.0350393446645251 85.14984083175659
0.031282561727695994 85.82515001296997
                                                                                                                           0.03438812578786855 84.0733597278595
0.03687326258255376 85.95911145210266
0.029598432248113333 86.35100626945496
0.028342962644442363 85.82388114929199
                                                                                                                          0.02342962644442363 85.82388114929199
0.04783162883379393 86.15980648994446
0.02354517739473118 85.59817576408386
0.03534271942313622 82.25598395927429
0.02087155810591799 82.59697151184082
0.021083436544156738 83.15034341812134
0.018854224610280385 85.5999858379364
0.01730320260308131 85.4431080703735
0.018228005781709075 86.89180874824524
0.017130475913829827 85.25211119651794
0.0163273202531311567 85.5942542552949
0.0159001647552941 87.77797293663025
                                                                                                                           8.1592736253151307 83.394234235294

87.77797293663025

8.169665317162967944 91.29987859725952

9.015067671697276333 86.02754735946655

9.015491769450751168 85.80844378471375
                                                                                                                           0.015491769450751168 85.80844378471375
0.017191762432318042 85.83576774597168
0.015099257135901737 86.19236254692078
0.013171155885275867 85.80043983459473
0.018089427783189065 82.20216989517212
0.013713264096252344 82.27552366256714
0.015097079953799645 82.61570501327515
0.014583172395187258 81.53680920600891
                                                                                                                           0.012903952738270164 82.076913356781
0.014689087841866745 81.84507703781128
0.013860434621434521 81.7358467578888
                                                                                                                           0.012902738625632116 82.06283044815063
0.013764038897567877 82.78730416297913
                                                                                   16
16
 [apurva.mandalika@g009\ code] \ \ python\ main.py\ \ -- task\ generate\ \ -- data\_split\ \ 'length'\ \ -- n\_layer\ 50
The file './tokenizer/length_vocab.json' exists. Loading tokenizer.
{'<pad>': 0, '<s>': 1, '</s>': 2, '<unk>': 3, 'I_NALK': 4, 'I_JUMP': 5, 'I_LOOK': 6, 'I_TURN_RIGHT': 7, 'I_RUN': 8, 'I_TURN_LEFT': 9, 'walk': 10, 'jump': 11, 'look': 12, 'turn': 13, 'right': 14, 'run': 15, 'left': 16, 'after': 17, 'twice': 18, 'opposite': 19, 'and': 20, 'thrice': 21, 'around': 22}
  loading model
  total params: 166848
```

| 3920/3920 [4:01:56<00:00, 3.70s/it]

Accuracy: 0.0000: 100%

Test accuracy: 0.0 [apurva.mandalika@g009 code]\$