A Multi-Perspective Analysis of Memorization in Large Language Models

Bowen Chen, Namgi Han, Yusuke Miyao
Department of Computer Science, The University of Tokyo
{bwchen, hng88, yusuke}@is.s.u-tokyo.ac.jp

Abstract

Large Language Models (LLMs) can generate the same sequences contained in the pre-train corpora, known as memorization. Previous research studied it at a macro level, leaving micro yet important questions under-explored, e.g., what makes sentences memorized, the dynamics when generating memorized sequence, its connection to unmemorized sequence, and its predictability. We answer the above questions by analyzing the relationship of memorization with outputs from LLM, namely, embeddings, probability distributions, and generated tokens. A memorization score is calculated as the overlap between generated tokens and actual continuations when the LLM is prompted with a context sequence from the pre-train corpora. Our findings reveal: (1) The inter-correlation between memorized/unmemorized sentences, model size, continuation size, and context size, as well as the transition dynamics between sentences of different memorization scores, (2) A sudden drop and increase in the frequency of input tokens when generating memorized/unmemorized sequences (*boundary effect*), (3) Cluster of sentences with different memorization scores in the embedding space, (4) An *inverse boundary effect* in the entropy of probability distributions for generated memorized/unmemorized sequences, (5) The predictability of memorization is related to model size and continuation length. In addition, we show a Transformer model trained by the hidden states of LLM can predict unmemorized tokens. [1]

1 Introduction

Large Language Models (LLMs), trained with enormous parameter and pre-train data sizes, like GPT-4 (OpenAI et al., 2024), show surprising performance on various tasks. Due to such enormous model size and pre-train data size, combined with the black-box nature of neural models (Alain and Bengio, 2018), LLMs present unique behaviors (Wei et al., 2022) that are unprecedented in previous machine learning, one of which is *memorization*.

Memorization (Hartmann et al., 2023) in the LLMs means the LLM can generate the same content recorded in their pre-train corpus. Being a coin with two sides, memorization in LLM can provide knowledge (Petroni et al., 2019) or cause personal information leakage (Yao et al., 2024). Previous research (Tirumala et al., 2022; Carlini et al., 2023; Biderman et al., 2023b) has studied memorization at the macro level, revealing memorization from the training phase or overlap between models, leaving more micro yet important questions left under-explored, e.g., what makes sentences memorized, dynamics of sentences with varying memorization scores transit to other scores when trained in a larger model, the dynamics when generating memorized/unmemorized sequence, its connection to unmemorized sequence, and its predictability.

To answer those questions, we study memorization from multiple perspectives. We form contexts of varying lengths from the pre-train corpora and input them into LLMs of different sizes. By controlling the maximum generated tokens, we collect the outputs (embeddings, decoding probabilities, and generated tokens) and compute the *memorization score* by comparing them with the actual continuations in the pre-train corpora. Our analysis connects memorization to model size, context size, continuation size, and unmemorized sequences. We also explored the dynamics of generating memorized/unmemorized tokens at input and output levels and the predictability of memorization. Our findings reveal:

(I) For both memorized or unmemorized sentences, their increase or decrease with model size is non-linear, indicating a maximum capacity for memorization. Memorized sentences decrease sub-

[1] The code of this study is at https://github.com/mynlp/memorizationstudy

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11190–11209
November 12-16, 2024 ©2024 Association for Computational Linguistics

linearly with continuation size and increase super-linearly with context size. We also analyzed the dynamics of how sentences with varying memorization scores transit to other scores when trained in a larger model. The results show that only limited low memorization score sequences will transit to higher scores when trained in larger models, and most memorized sentences are also inherited when trained in larger models.

(II) A *boundary effect* was observed when the model began generating memorized/unmemorized sequences. The n-gram frequency suddenly decreases when generating unmemorized tokens and suddenly increases when generating memorized tokens, which is also observed at the sequence level. The significance of the boundary effect for memorized sentences decreases with the increase in model size. However, as large models contain more memorized sequences, this suggests large models have a lower threshold in the value of the boundary effect to memorize a sentence. This indicates the value of the boundary effect relates to the difficulty of memorizing a sentence.

(III) The embedding dynamics analysis showed sentences with different memorization scores cluster in the embedding space, where the mutual embedding distance grows with model size. Sentences of close memorization scores are also close in embedding space, suggesting the existence of paraphrase memorization.

(IV) We analyzed decoding dynamics by examining entropy over vocabulary and the drift of decoded embeddings. Entropy analysis revealed an *inverse boundary effect*, where entropy suddenly increases for unmemorized sequences and decreases for memorized sequences.

(V) We trained a Transformer model to discuss the predictability of memorization. The results suggest that predicting memorization is easier in large models and easier when predicting unmemorized sequences, which can be explained by the significance of the boundary effect.

2 Related Works

2.1 Scaling Laws of LLM

In this study, our experiments span across various model sizes. This relates to the research of Scaling Laws (Kaplan et al., 2020; Abnar et al., 2021; Villalobos, 2023), which suggests the performance of LLM scales with the corpora size, the parameter size, and the computation required.

Inspired by the Scaling Law, researchers scaled the LLMs in both model and data size to gain higher performance like T5 (Raffel et al., 2023), GPT-3 (Brown et al., 2020), PaLM 2 (Anil et al., 2023). On the other hand, researchers also analyzed how scaling affects particular tasks like translation and prompt injection attack (Sun and Miceli-Barone, 2024). Within those discussions, the emergent abilities of LLM (Wei et al., 2022) are discovered. This means the LLM suddenly reaches high performance on previous low-performance tasks when reaching a certain model size. Recent studies have questioned whether emergent abilities are mirages caused by metrics misuse (Schaeffer et al., 2023) or just context-learning (Lu et al., 2023).

Regarding scaling in the field of memorization, Carlini et al. (2023) discussed the memorization across model size and found that the number of memorized texts grows with the model size and context size. Additionally, Biderman et al. (2023a) also discussed similar topics and found that a large portion of memorized text in a small-size model is also memorized by a larger model, showing that memorized texts may share common features.

2.2 Memorization

Prior to LLM, over-fitting is close to memorization (Tetko et al., 1995), which means a near-zero loss in the train set, suggesting the model memorized the input and its label (Zhang et al., 2017). However, memorization differs from overfitting because LLMs can perform well on the test set, whereas overfitting usually results in poor test set performance. Feldman (2020) analyzed the necessity of memorization in classification models. They demonstrated that in a long-tail distribution, where many categories have only a few samples, the neural model struggles to extract general features. Instead, the best strategy for the neural model is simply to memorize these samples and compare them with the data in the test set.

LLMs, unlike classification models, can directly generate their pre-train content, making the memorization observable. This can be used to form knowledge graphs (Petroni et al., 2019; AlKhamissi et al., 2022), but also leads to data contamination (Sainz et al., 2023) and privacy risk (Yao et al., 2024). Previous research has discussed memorization on a macro level. Tirumala et al. (2022) showed large models tend to memorize samples more easily in training. Carlini et al. (2023) dis-

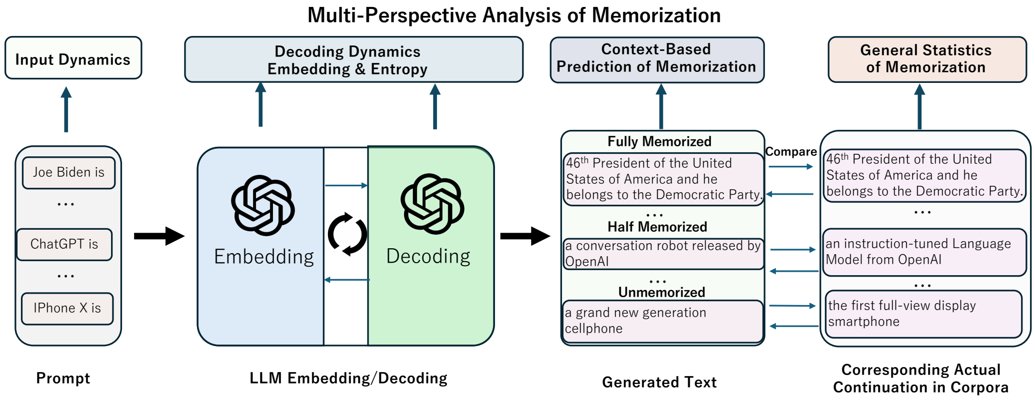

Figure 1: Memorization and research scope in this study. We prompt a partial context text into the model and calculate the memorization score of the generated continuations over the whole corpora. We show how inputs, model factors, and generation dynamics affect the memorization results.

cussed LLM memorization from factors like model size, continuation size, and context size, showing their inter-correlations. Biderman et al. (2023a) studied memorization in LLMs in their training phase and the overlap between different model sizes, finding commonly memorized sentences. Li et al. (2024) studied memorization by using hidden states in LLM in several specified datasets. Prashanth et al. (2024) set a threshold of a number of repetitions in the pre-train corpora to separate memorized sequences and predictable sequences. Speicher et al. (2024) created a sandbox setting to exploit the memorization of random strings.

Previous research discussed from a macro-level where they analyzed the memorization from an overall perspective, focusing mostly on fully memorized sequences with less consideration of the transition between sentences and how it is related to the model's inner-working mechanism. This study focuses on micro yet under-explored questions, e.g., the dynamics while generating memorized sequences, why some sentences are memorized, the relation to unmemorized sequences, and its predictability.

3 Experiment Setting

3.1 Experiment Overview

As shown in Figure 1, we collect LLM outputs (tokens, probability distributions, and embeddings) and calculate a memorization score for every sentence to evaluate its extent of memorization and obtain general statistics. Then, we conduct an in-depth analysis of these collected outputs, examining the dynamics of both the prompted context input and the generated tokens. Finally, we discuss the predictability of memorization.

3.2 Memorization Score

We prompt the LLM with context sequence tokens $C = \{c_1 \cdots c_l\}$ from its pre-train corpora and use greedy decoding to generate the predicted continuation tokens $X = \{x_1 \cdots x_n\}$. We also collect the actual continuations $Y = \{y_1 \cdots y_n\}$ under this context. This process is iterated for the whole corpora. The memorization score is calculated as follows:

$$M(X, Y) = \frac{\sum_{i=1}^{n} \mathbf{I}(x_i = y_i)}{n} \quad (1)$$

n means the length of the continuation tokens. $\mathbf{I}$ is the Indicator Function. If $M(X, Y) = 1$, the sequence is fully memorized under this context sequence, termed *K-extractable* (Carlini et al., 2021). A sequence Y is unmemorized if $M(X, Y) = 0$.[2]

3.3 Criteria for Memorization Prediction

We use Token Accuracy and Full Accuracy to evaluate performance in predicting memorization.

The prediction for a sequence of generated tokens is $\hat{X} = \{\hat{x_1}, \hat{x_1} \ldots \hat{x_n}\}$ where each prediction $\hat{x_i}$ is a binary label indicating the token at this index is memorized or not. The gold label is denoted as $\hat{Y} = \{\hat{y_1}, \hat{y_2} \ldots \hat{y_n}\}$ where each $\hat{y_i}$ is the golden label. The Token Accuracy is defined as:

$$T(\hat{X}, \hat{Y}) = \frac{\sum_{i=1}^{n} \mathbf{I}(\hat{x_i} = \hat{y_i})}{n} \quad (2)$$

A prediction for a sequence is defined as fully correct if $T(\hat{X}, \hat{Y}) = 1$. We obtain Full Accuracy by

[2] Based on the different number of continuation tokens, the memorization score has different granularities. In some experiments, we classify sentences into several memorization levels based on their memorization scores.

dividing the number of fully correct sequences by the number of all sequences.

3.4 Model Setting

We use the Pythia (Biderman et al., 2023b) to analyze the memorization as it provides LLMs trained across various sizes with the same training order using open-sourced Pile (Gao et al., 2020) corpora that ensure experimental stability. We investigate the model size of [70m, 160m, 410m, 1b, 2.8b, 6.9b, 12b] where m and b stands for million and billion. We choose the LLM trained on the deduplicated Pile corpora to avoid the effect of duplicated sentences, as previous research reported the chance to be memorized grows exponentially with the number of duplicates (Kandpal et al., 2022).[3]

3.5 Corpora Setting

The open-sourced Pile (Gao et al., 2020) corpora are publicly available data. [4] The data contains 146432000 rows with a chunk length of 2048, reaching a data size of around 800GBs. The experiment is iterated through the whole training data matrix, meaning that we did not conduct sampling over the rows. Instead, we iterated the whole Pile matrices. For example, if the context size is 32 and the continuation size is 96, we prompt the first 32 tokens at each row into the model. We use the Pythia model to generate the 96 tokens equal to the continuation size. Then, we compare the generated token IDs with the gold token IDs in the data to calculate the memorization score at each row. This process is repeated for the entire Pile matrix, distributed over different CUDA devices.

4 Experiment Results

4.1 Memorization Factors

This section discusses how memorization is connected to model size, context size, and continuation size. We collect the number of sentences with different memorization scores under different model sizes. Then, we divide those sentences with different memorization scores into ten sets with a memorization score difference of 0.1, as shown in Figure 2.

[3]For more details, please refer to https://github.com/EleutherAI/pythia

[4]The used data can be downloaded at https://huggingface.co/datasets/EleutherAI/pile-deduped-pythia-preshuffled/tree/main

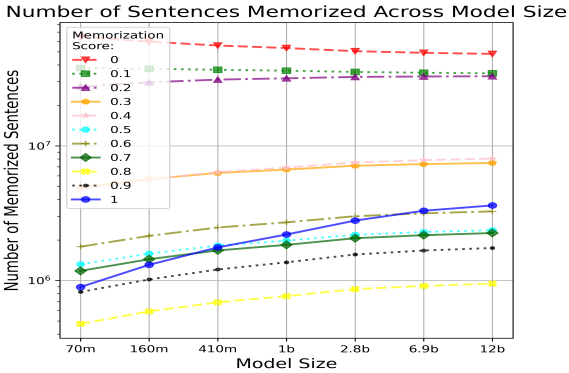

Figure 2: Number of sequences of different memorization scores across different model sizes.

Firstly, we can see that the memorized sentences increase with the model size and context size but decrease with the increase of continuation size, which aligns with previous research (Carlini et al., 2023). We illustrate a more in-depth analysis in the following sections.

4.1.1 The Factor of Model Size

In this experiment, we discuss how model size affects the number of memorized and unmemorized sentences. From Figure 2, we can obtain that:

(I) The number of sentences with low memorization scores (0-0.3) is significantly higher than those with high memorization scores, indicating that most of the pre-train data are not memorized despite the existence of memorization in LLMs.

(II) Among sentences with high memorization scores, the count of fully memorized sentences increases more rapidly, suggesting LLMs tend to memorize sentences fully rather than partially. Additionally, the number of sentences with low memorization scores (0, 0.1) decreases as the model size increases, indicating that unmemorized sentences gradually become memorized with larger models.

(III) The increase or decrease in the number of memorized or unmemorized sentences is not linear with respect to model size. There is a noticeable increase in numbers for fully memorized sentences from 70 million to 2.8 billion parameters, compared to 2.8 billion to 12 billion parameters. A similar decreasing trend for unmemorized sentences is observed. This suggests a capacity for memorization, implying LLMs cannot memorize the entire corpus even with sufficiently large model sizes.

4.1.2 Context and Continuation Size

We change the length of the context and fix the length of continuations or vice versa. Then, we ob-

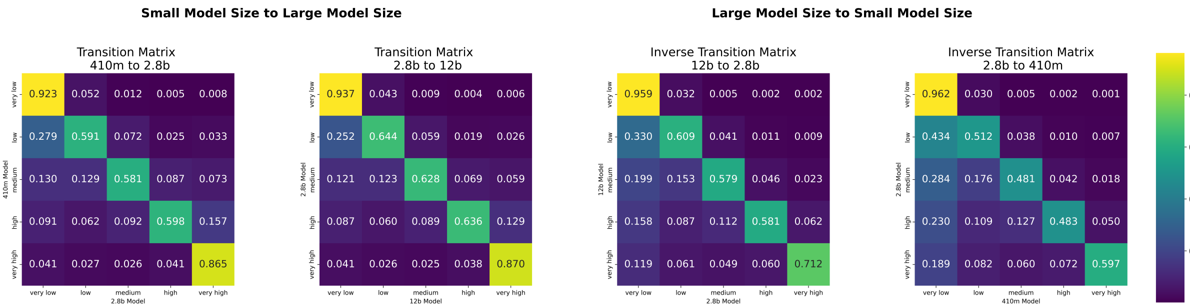

Figure 3: Transition matrix of sentences with different memorization scores. The two left figures are the transition matrix from small size model to large size, with the left one being the transition matrix from 410m to 2.8b and the right one being 2.8b to 12b. The two right figures are the inverse transition matrix from large size model to small size model, with the left one being 12b to 2.8b and the right one being 2.8b to 410m.

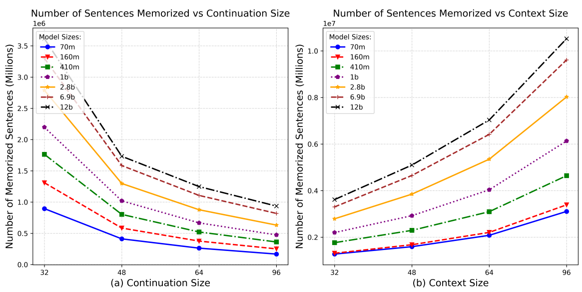

Figure 4: Number of memorized sentences in different continuation and context sizes across model size.

serve the change in the number of fully memorized sequences shown in Figure 4 (a) and (b). We can conclude:

(I) The decrease of memorized sentences with increasing continuation size is not linear. For instance, the continuation size increase from 64 to 96 results in a relatively minor decrease compared to the change from 32 to 48, indicating some sentences are firmly memorized.

(II) The reduction in memorized sentences with increased continuation size is more obvious in larger models. This demonstrates that although larger models memorize more sentences, most of their memorized sequences are less rooted compared to those of smaller models.

(III) The increase in memorized sentences with context size is also non-linear, with longer context leading to an almost exponential rise in the number of memorized sentences. This increase is more significant in larger models, indicating more sequences are potentially memorized in large models, which can be elicited by giving longer context.

4.2 Memorization Transition

This section discusses how sentences with different memorization scores transit across model sizes. Specifically, when trained with a larger model, we study the mutual transition between sentences with different memorization scores. We classify sentences with memorization scores with a range difference of 0.2 with labels of very low, low, medium, high, and very high. We plot the transition matrices in Figure 3, which shows the transition from the 410m size model to the 2.8b size and 2.8b size to the 12b size model and their inverse transition matrices. We can conclude:

(I) Most sentences remain in their previous state even when trained with a larger model, as indicated by the diagonal entries in the transition matrices. Additionally, the higher the memorization score, the less likely the sentence will transit to another state. For highly memorized sentences, over 90% remain memorized compared to those with low memorization scores. In the inversed transition matrix of the large-to-small size model, we see that most sequences also tend to stay in their original state with probability transferring to lower memorization score states, which fits our expectations.

(II) With increasing model size, sentences are more likely to stay at their original memorization level. For example, the transition probability at the diagonal is higher in the transition matrix from 2.8b to 12b compared to that of 410m to 2.8b. This suggests that memorized or unmemorized sequences become fixed as the model size increases. For the inversed matrix, the model is more likely to transfer to a lower memorization score state. In particular, transferring to the very low state is more probable than other low memorization score states. This shows when LLM starts to forget a sequence, it

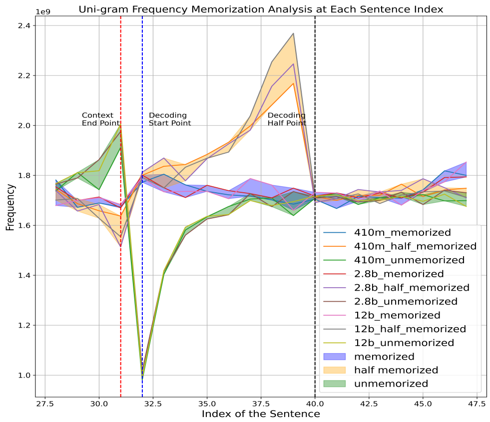

Figure 5: Uni-gram frequency at each index for memorized, unmemorized, and half-memorized sequences. The context length is 32. The continuation length is 16. We label the end of context, the beginning of encoding, and half of the encoding with red, blue, and black lines.

tends to forget that sequence completely.

(III) For high memorization score sentences, there is only a very small chance to transit to a low memorized state, implying most sentences memorized by small models are also inherited by the larger ones. This shows the memorized sentences are not memorized randomly but share certain common features with a little share of randomness. This is also observable in the inverse transition matrix, where there is little chance of transferring to a high memorization score state even when the model size decreases.

4.3 Frequency Dynamics of Input Sequences

In this section, we discuss the question of *whether there is any sign when the model starts to generate memorized or unmemorized sentences.* Especially what makes sentences memorized at different extents and why some sentences are memorized by large models but not small models.

4.3.1 Token Level Frequency Analysis

Firstly, we begin with the input level by analyzing the frequency of the n-grams in the pre-train corpora. We show the uni-gram frequency across steps in Figure 5 [5], and we can see:

(I) A clear *boundary effect* is observed around index 32, representing the first generated token. The frequency drops and then rises for memorized sentences (*positive boundary effect*), whereas it rises and then drops for unmemorized sentences (*negative boundary effect*). The negative boundary effect is more significant than the positive one.

(II) For the half-memorized sentence, we see the frequency increases and drops (*negative boundary effect*) around the index of 39, which is half the length of generated tokens. This shows that for the half-memorized sentence, the previous half is mostly memorized, and the later half is mostly unmemorized, meaning that the memorized tokens are distributed in a near continuous way rather than scattered in the generated tokens.

(III) The positive boundary effect in memorized sentences suggests that memorization is driven by the higher frequency of initial tokens, implying that remembering the first few tokens makes the entire sentence easier to memorize. Conversely, the negative boundary effect in unmemorized sentences indicates that the low frequency of initial tokens makes the following sequences easier to forget.

4.3.2 Sequence Level Frequency Analysis

Given the existence of the token level boundary effect, we extend the discussion to the sequence level. As shown in Table 1, we calculate the average n-gram frequency of context and continuation and boundary frequency difference. We can obtain:

(I) The frequency of uni-grams is significantly higher than that of bi-grams, approximately 3.5 times higher on average in both context and continuation. The boundary effect is consistent in the bi-gram setting, though the positive boundary effect is less obvious due to the frequency drop when computed with bi-grams. Despite this, the actual frequency gap remains substantial (million level) when considering the unit is billion.

(II) In memorized sentences, the frequency is lower in the context and higher in the continuation of the memorized data (M column), whereas unmemorized sentences exhibit the opposite pattern. This suggests the boundary effect also exists at the sequence level, though less obvious.

(III) For half-memorized sentences, the frequency of continuation tokens is higher than the context average frequency. This is due to the frequency increase before reaching the first generated unmemorized tokens, as indicated in Figure 5.

(IV) In the Boundary Frequency Difference column, both positive and negative boundary effects decrease with increased model size. However, the decrease in the positive boundary effect makes it less significant, while the decrease in the negative

[5] N-gram statistics overall steps are in the A.5.

Size	Average Context Frequency						Average Continuation Frequency						Bounrady Frequency Difference					
	Uni-gram			Bi-gram			Uni-gram			Bi-gram			Uni-gram			Bi-gram		
	M	H	U	M	H	U	M	H	U	M	H	U	M	H	U	M	H	U
160m	1.708	1.713	1.744	0.551	0.534	0.691	1.739	1.837	1.628	0.535	0.659	0.567	0.114	0.330	-0.939	0.033	0.101	-0.663
1b	1.713	1.711	1.752	0.558	0.552	0.697	1.736	1.832	1.631	0.509	0.682	0.564	0.103	0.270	-0.981	0.028	0.090	-0.696
6.8b	1.721	1.710	1.759	0.570	0.565	0.701	1.736	1.829	1.638	0.496	0.699	0.564	0.090	0.140	-0.963	0.027	0.085	-0.726
12b	1.721	1.720	1.760	0.572	0.569	0.702	1.736	1.846	1.626	0.493	0.704	0.563	0.039	0.237	-1.016	0.026	0.083	-0.732

Table 1: Uni-gram and bi-gram statistics. The frequency unit is billion. Boundary Freq Difference means we use the first generated token's frequency to subtract the last token's frequency in the context (i.e., the boundary effect). M, H, and U mean memorized, half-memorized, and unmemorized, respectively.

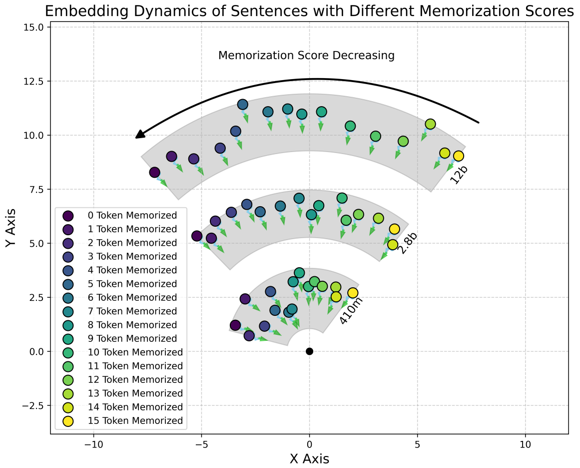

Figure 6: Embedding dynamics across 410m, 2.8b, and 12b model sizes. x Token Memorized means x generated tokens are the same as the true continuation. The arrow for each cluster means the visualized moving direction of generated tokens. The gray area means the span of those clusters.

boundary effect makes it more significant. This implies the significance of the positive boundary effect correlates with the ease of memorizing a sentence. In contrast, the significance of the negative boundary effect correlates with the difficulty of not memorizing a sentence.

4.4 Decoding Dynamics

In this section, we analyze the dynamics of the output, e.g., the movement of generated embeddings at each step and corresponding entropy changes for sequences of varying memorization scores.

4.4.1 Embedding Dynamics

We analyze embeddings of generated tokens for sentences with varying memorization scores. We collect the hidden state of the last layer for each generated token for sentences with different memorization scores and compute the pair-wise Euclidean distance and cosine similarity. We draw the following figure to represent the embedding dynamics shown in those pair-wise Euclidean distance and cosine similarity. [6] We can obtain:

(I) The mutual angle remains stable across different encoding steps between sentences of different memorization scores, which also suggests a stable mutual cosine similarity. Meanwhile, since they have the trend of moving toward the center, the Euclidean distance also decreases with the generation of tokens.

(II) Sentences with high memorization scores are close in the embedding space. This suggests their generated sequences are both semantically and lexically similar to the actual continuation. This indicates the existence of paraphrase memorization since those high memorization score sequences are probably only different in a few tokens while sharing the same meaning.

(III) Larger models exhibit larger mutual Euclidean distances and angles. The increase in angle leads to a decrease in cosine similarity. The reason can be attributed to the expansion of hidden sizes (e.g., 512 for 70m, 2048 for 1b model). The expansion of hidden size increases the expressivity of the embedding and enlarges the mutual distances, making different sentences more differentiable. This also helps to explain the performance gap between different model sizes: larger models distribute different sentences more distinctly with fewer embedding overlaps, while smaller models mix embeddings more, leading to ambiguity and degraded performance.

4.4.2 Generation Dynamics and Entropy

This section discusses the generation dynamics when the LLM generates sentences with different

[6] Detailed numbers of Cosine Similarity and Euclidean distance between sentences with different memorization scores at different decoding steps are at Appendix A.3

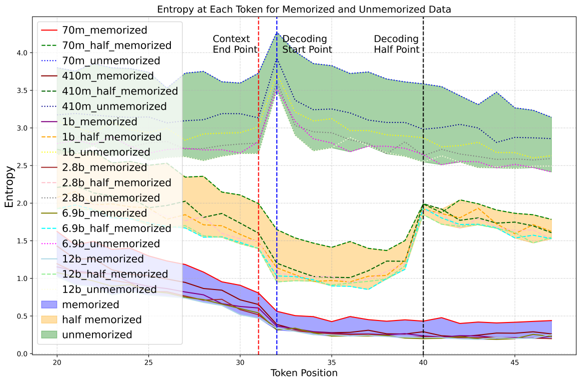

Figure 7: Averaged entropy dynamics at each index for memorized, unmemorized, and half-memorized sequences. The prompted context length is 32, and the continuation length is 16. We label the end of context, the beginning of encoding, and the half of the encoding with red, blue, and black lines.

memorization scores. Given the boundary effect at the input frequency, we ask *whether the model exhibits similar behaviors when generating memorized or unmemorized sequences.* As shown in Figure 7, we collect the probability distributions for each generated token of memorized, unmemorized, and half-memorized sentences, with 10,000 samples for each, and calculate the average entropy at each token. We can conclude:

(I) First, we can see that the entropy differs based on memorization scores and model size. Unmemorized sentences have a higher average entropy at each token than memorized sentences. This shows LLM is more confident when generating memorized sequences.

(II) Additionally, the entropy drops when generating memorized sequences (fully memorized and former half of half-memorized sequences) and increases when generating unmemorized ones (unmemorized and later half of half-memorized sequences). This shows an *inverse boundary effect* in entropy compared to the one in frequency. However, the entropy drop for memorized sentences is not samely significant as the half-memorized sentence, as LLM is more confident when generating memorized sequences, which is naturally low entropy, leaving less space for the entropy drop.

(III) We can also see the entropy decreases with the increase in model size. This suggests that larger models are more confident about generating tokens than small ones. Additionally, the entropy of context for memorized tokens is also lower, showing that the entropy of context also relates to whether its continuations are memorized. Further, the significance of the inverse boundary effect in entropy decreases with the model size for memorized sequences while remaining unchanged for unmemorized sequences. This differs from the boundary effect in frequency, whose significance decreases with increased model size for memorized sequences while increasing for unmemorized ones.

4.5 Prediction of Memorization

Since similar context can trigger memorized texts (Stoehr et al., 2024), and the generated token may also be a paraphrase of the actual continuation (Ippolito et al., 2023), common methods that check memorization by searching the generated tokens in the corpora are challenged. Thus, it would be beneficial if it were possible to predict the memorization by embedding. We sample sentences with different memorization scores evenly from the whole corpora. A Transformer (Vaswani et al., 2023) is trained to predict a binary label at each continuation token, predicting whether this token is memorized by receiving all embeddings generated so far and related statistics, e.g., the entropy and frequency.[7]

4.5.1 Results on Prediction of Memorization

We discuss the predictability of predicting memorization based on results presented in Table 2. We can obtain:

(I) First, regarding Token Accuracy, we can see with a naive Transformer model, the token-level accuracy can reach 80% accuracy or even higher, showing that the prediction of memorization at the token level is easy. The Full Accuracy is low as this requires a correct prediction for every token.

(II) In either continuation length, both token accuracy and full accuracy increase with model size. This shows the prediction of memorization is easier for large models because the greater embedding distances make classification easier.

(III) Token Accuracy increases with the continuation size in either model size, likely due to increased training data. For instance, continuation length 64 contains four times larger tokens than that of continuation length 16. However, Full Accuracy decreases as the continuation length increases as more tokens need to be predicted correctly.

4.5.2 Analysis of Full Accuracy

In this experiment, we analyze how fully correct predictions are distributed across memorization

[7]Experiment details and settings are in Appendix A.1.3.

Len	70m		410m		1b		2.8b		6.9b		12b		Dist	
	Token	Full	Token	Full	Token	Full	Token	Full	Token	Full	Token	Full	M	U
16	78.2	10.2	78.6	10.4	78.8	10.6	80.1	10.7	77.4	8.3	80.3	**10.9**	53.1	46.9
32	78.6	5.9	79.6	6.0	79.7	6.1	80.1	6.3	80.5	6.4	80.8	6.4	51.6	48.4
48	79.6	5.2	80.3	5.4	80.4	5.6	80.4	5.5	80.8	5.8	81.0	6.0	51.0	49.0
64	80.1	4.7	80.8	4.8	81.2	5.2	81.5	5.5	81.8	5.8	**82.1**	6.0	50.7	49.3

Table 2: Performance on memorization prediction at context length 32. Len means the length of continuation tokens for prediction. Token and Full mean Token and Full Accuracy. The various model sizes in the column show which size of LLM's embeddings are used to train the Transformer. The M and U in the Dist column mean the distribution of memorized and unmemorized tokens used to train and evaluate. For each continuation length, the best results across model sizes are in bold. The best results in a model size across continuation lengths are underlined.

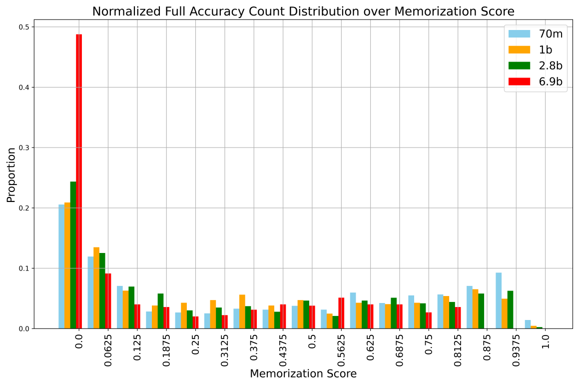

Figure 8: Full accurate predictions distribution at different memorization scores across model size

scores to discuss whether the difficulty in predicting memorization changes with the memorization score as shown in Figure 8, and we can obtain:

(I) The Transformer model trained with embeddings from any LLM size is better at predicting sentences with low memorization scores, even when the label distribution is close. The low portion of sentences with high memorization scores indicates they are harder to predict accurately.

(II) As model size increases, the proportion of low memorization scores rises and decreases for high memorization score sentences, which even reaches zero for the 6.9b model. This suggests predicting unmemorized sentences is easier in large models compared to memorized ones.

(III) A possible explanation for the above behaviors can be made. Previous experiments show that the boundary effect of unmemorized sequences is more significant than that of memorized sequences in both token frequency and entropy. Additionally, the significance of the boundary effect decreases for memorized sequences while increasing for unmemorized sequences with increased model size. With the decrease of significance in such features for memorized sequences, they become hard to predict. The unmemorized sequences become easy to predict as the significance increases.

4.6 Conclusion

In this study, we comprehensively examined LLM memorization from various perspectives. At the statistical level, we extended previous research to include sentences with lower memorization scores and conducted experiments showing memorization transitions across model sizes. Analyzing input dynamics through frequency analysis, we identified positive and negative boundary effects when generating memorized and unmemorized tokens, indicating their relation to the ease of memorization. In the output dynamics, at the embedding level, we found clusters of sentences with different memorization scores in the embedding space, and the close distance of sentences with high memorization scores indicates the existence of paraphrase memorization. At the entropy in the output dynamics, we observed an inverse boundary effect and analyzed its change with model size. Finally, we trained a Transformer model to predict memorization, showing that token-level prediction is easy while sentence-level is challenging. Through analysis of fully correct predicted samples, we found unmemorized tokens are easier to predict than memorized tokens, which can be explained by the significance of the boundary effect.

5 Acknowledgments

We sincerely thank every reviewer for their valuable suggestions. This work is funded by the Institute of AI and Beyond of the University of Tokyo.

6 Limitations

This research has analyzed various factors regarding the memorization behavior of LLMs. We acknowledge that there are still limitations to this research. Due to the lack of LLMs whose models and data are both being released, it is hard to compare the memorization across different LLMs since even if they were released, the pre-train data differs based on different LLMs. Future research can focus on how to comprehensively measure the memorization of various LLMs, either close-sourced like GPT-4 or open-sourced like Pythia. Additionally, Pythia only provides LLMs up to 12b, which is still considerably small when compared to SotA open-sourced LLMs like LLaMa2, whose largest model is 70b. The emergent abilities reaching 70b size may also affect memorization. Though we trained a Transformer model to predict the memorization of LLMs, it is more analysis-oriented, proving the possibility of predicting memorization. Thus, the performance in the prediction experiment is not the main focus.

Additionally, in this research, the discussion of memorization is under the context when the LLM generates tokens that are the same as the actual continuations in the corpora, defined as *K-extracble* in this study. It is possible that paraphrase memorization exists. However, it is difficult to identify such behavior on a large scale, especially across whole corpora. This is because the identification of paraphrase memorization can not be simply decided by either token overlap or embedding similarity since neither of them can truly compare whether a sentence is a paraphrase. This means we have to identify by more complex methods, e.g., human annotators or a trained classification model. However, a classification model cannot be 100% accurate when applied to the corpora level, leading to false positives and false negatives that influence the analysis results. However, paraphrase memorization can be left to future research.

7 Ethical Considerations

In this study, we prompted the Pythia model with context tokens from its pre-train corpora. The data released by Pythia are all tokenized and turned into token IDs, so the original information is not visible. We have provided a case study of the prediction of memorization in the Appendix A.2 by turning the token IDs into texts. However, we did not observe any personal information or offensive language during this process. Additionally, we obeyed related open-source licenses of both Pythia and Pile corpora. Thus, this study is not concerned with ethical issues.

References

Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. 2021. Exploring the limits of large scale pre-training.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023a. Emergent and predictable memorization in large language models.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023b. Pythia: A suite for analyzing large language models across training and scaling.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Vitaly Feldman. 2020. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, page 954–959, New York, NY, USA. Association for Computing Machinery.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models.

Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. 2023. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia. Association for Computational Linguistics.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

Bo Li, Qinghua Zhao, and Lijie Wen. 2024. Rome: Memorization insights from text, logits and representation.

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2023. Are emergent abilities in large language models just in-context learning?

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language models as knowledge bases?

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V au2, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. 2024. Recite, reconstruct, recollect: Memorization in lms as a multifaceted phenomenon.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*.

Till Speicher, Aflah Mohammad Khan, Qinyuan Wu, Vedant Nanda, Soumi Das, Bishwamittra Ghosh, Krishna P. Gummadi, and Evimaria Terzi. 2024. Understanding the mechanics and dynamics of memorisation in large language models: A case study with random strings.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models.

Zhifan Sun and Antonio Valerio Miceli-Barone. 2024. Scaling behavior of machine translation with large language models under prompt injection attacks. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 9–23, St. Julian's, Malta. Association for Computational Linguistics.

Igor V. Tetko, David J. Livingstone, and Alexander I. Luik. 1995. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of Chemical Information and Modeling*, 35(5):826–833.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Pablo Villalobos. 2023. Scaling laws literature review. Accessed: 2024-04-29.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization.

A Appendix

A.1 Experiment Setting

A.1.1 Pythia LLM Generation

The experiment uses 64 A100 40Gbs GPUs when using LLMs to generate tokens given the previous context, which utilizes PyTorch's parallel running packages. We run the model with half-precision, which increases both the speed and saves memory. This follows the previous Pythia implementation when generating tokens given context.

The running time depends on the model size and the generated token length. When a 70m model with 32 context tokens and 16 tokens is required to be generated, it can be run with one A100 GPU within several hours. However, if such an experiment is in a single A100 GPU, it would estimated to take two weeks to finish the generation. Therefore, with 64 GPUs, the running of the 12b model in such a situation can be shortened to around one day. However, the generation time also largely increases with the length of generated tokens, which grows linearly with the number of tokens to be generated. Additionally, since we use greedy decoding and do not consider other possible decoding options but the token with the highest probability, it is also possible that the running time may be different if using a more complicated decoding strategy. However, using a more complex decoding strategy does not affect the results. If a sequence of tokens is memorized, the memorized token will always be the most probable token when generating them.

A.1.2 N-Gram Statistics Caculation

For the analysis of 4.3, since calculating n-gram statistics for pre-trin corpora demands lots of memory, we used another experimental environment with 128G RAM. It takes several days to calculate each gram's statistics. Also, storing the result of n-gram statistics takes over 1TB of storage.

A.1.3 Memorization Prediction

Regarding the prediction of memorization, we sample 2,000 sentences at the sentence set of each memorization score. For example, in the situation where the continuation length is 16, we will sample 2,000 sentences from memorization scores, the range of which is from 0 to 1 with 0.0625 as the unit. The Transformer model receives embedding of the last layer at each generated step with the corresponding entropy and uni-gran frequency and then outputs an embedding at each step. This output embedding from the Transformer model will be used to pass through a linear layer, and the Softmax function calculates the probability that the token at this index is a memorized token or not. The training is conducted using a 4-layer Transformer model with a dropout probability of 0.5, and the learning rate is 1e-4. Additionally, the only changing parameter with the increase in model size is the hidden size, as the larger model has a larger hidden size. The training and evaluation are conducted across five random seeds, and we report the average performance. The train, valid, and test split ratio is 0.6, 0.2 and 0.2. Additionally, we make sure that the distribution of sentences of different memorization scores is even in the dataset; thus, the model does not make biased predictions.

A.2 Case Study

We also provide a case study of the model's prediction on the test set regarding the prediction of memorization in Figure 9. From this figure, we can see that:

1. Confirming previous experiments, the memorized token is mostly continuous, showing the memorization happens in chunks of sequences rather than individual sequences.
2. In the first example, the model outputs a fully correct prediction that aligns with the actual label, showing the possibility of predicting memorization by utilizing embedding information.
3. In the second example, we see that the model's prediction does not align with the actual continuation. The model predicts an unmemorization label for the memorized label.
4. In the third example, we can see this unmemorized sequence. The model fully predicted those labels. We also noticed that the probability for the corresponding token is very high, showing that the model is very confident about the prediction.

A.3 Supplementary Data for Embedding Dynamics

We have presented the PCA visualized embedding dynamics in Figure 6. In this section, we provide the actual numbers of both Cosine Similarity and Euclidean distance to further illustrate this point.

Similarly to Figure 6, we also provide detailed cosine similarity and distance results from Figure 10 to Figure 18. From those figures, we can see that the cosine similarity fluctuates but remains relatively stable for sentences with different memorization scores. Additionally, sentences that are

Context	Label		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
How many minutes are there between 2:02 AM and 2:01 PM? 7	Text		19	\n	how	many	minutes	are	there	between	5	:	46	PM	and	3	:	32
	Pred Prob		0.90	0.99	0.94	0.93	0.83	0.95	0.89	0.95	0.95	0.94	0.98	0.54	0.83	0.93	0.91	0.97
	Gold	0.69	M	M	M	M	M	M	M	M	U	M	U	U	M	U	M	U
	Pred	0.69	M	M	M	M	M	M	M	M	U	M	U	U	M	U	M	U
The objective of all equity funds is to seek out profit opportunities. Types of equity	Text		funds	\n	\n	There	are	different	types	of	equity	funds	,	categorized	according	to	risk	levels
	Pred Prob		0.97	0.75	0.60	0.83	0.89	0.93	0.98	0.97	0.99	0.99	0.98	0.97	0.98	0.94	0.99	0.99
	Gold	0.65	M	M	M	M	M	U	M	M	M	M	U	U	U	U	U	U
	Pred	0.5	U	M	U	U	U	U	U	U	U	U	U	U	U	U	U	U
Creating men whose expectations of what they should look like are unattainable. In recent years,	Text		researcher	at	the	University	of	California	School	of	Medicine	injected	the	brain	of	diabetic	rats	with
	Pred Prob		0.93	0.95	0.80	0.99	0,99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
	Gold	0	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U
	Pred	0	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U	U

Figure 9: Prediction Examples. Pred Prob means the output predicted probability of the corresponding label in the pred row for each example. Gold means the true label. Pred means the predicted label, and Pred Prob means the probability of the corresponding prediction. M means the label for a token at this index is a memorized token. U means the label for a token at this index is an unmemorized token.

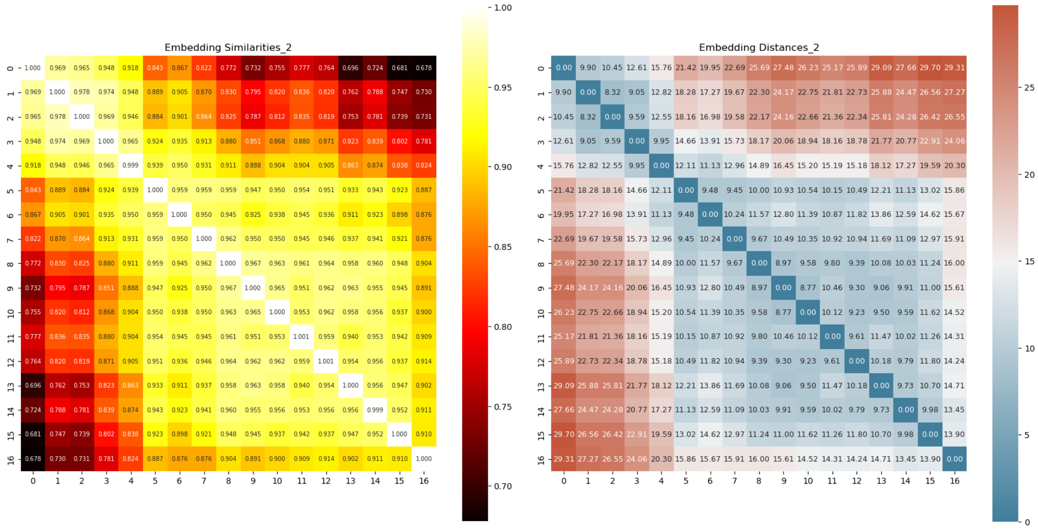

Figure 10: 410m Model, Step 2

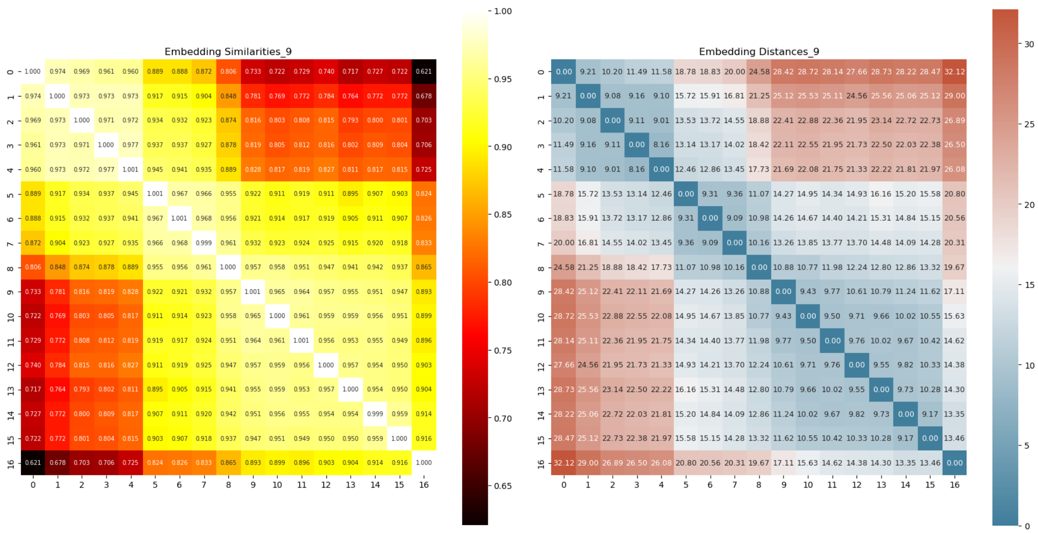

Figure 11: 410m Model, Step 9

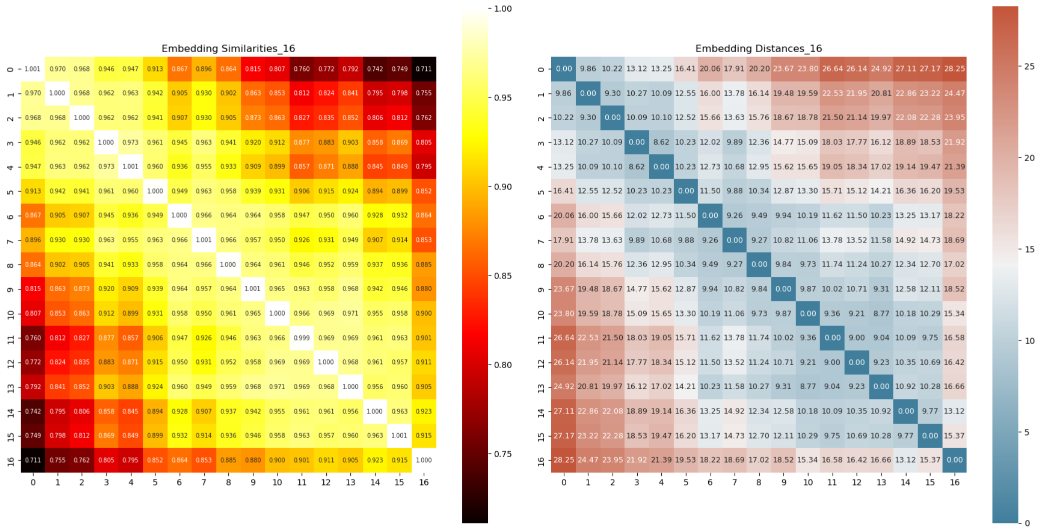

Figure 12: 410m Model, Step 16

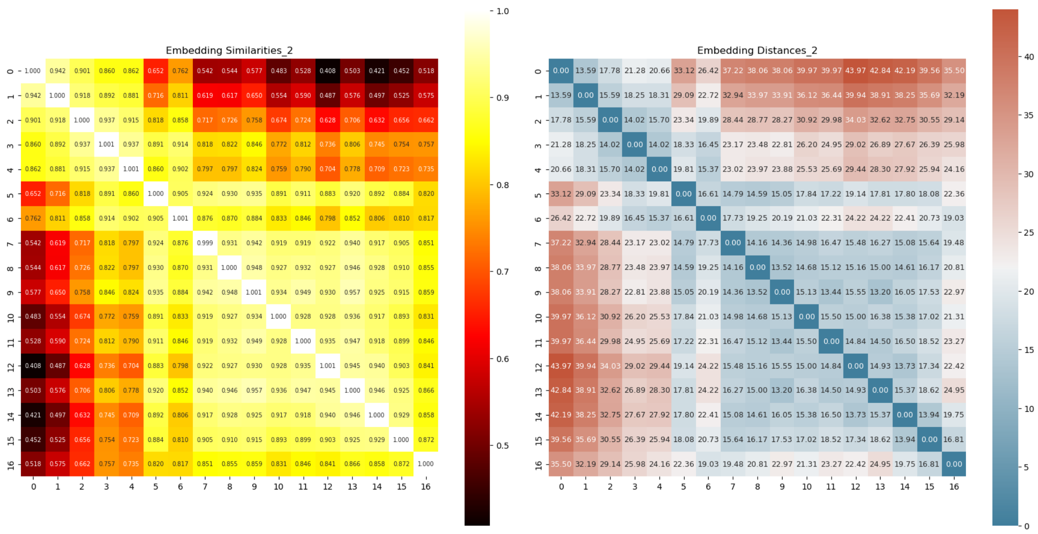

Figure 13: 2.8b Model, Step 2

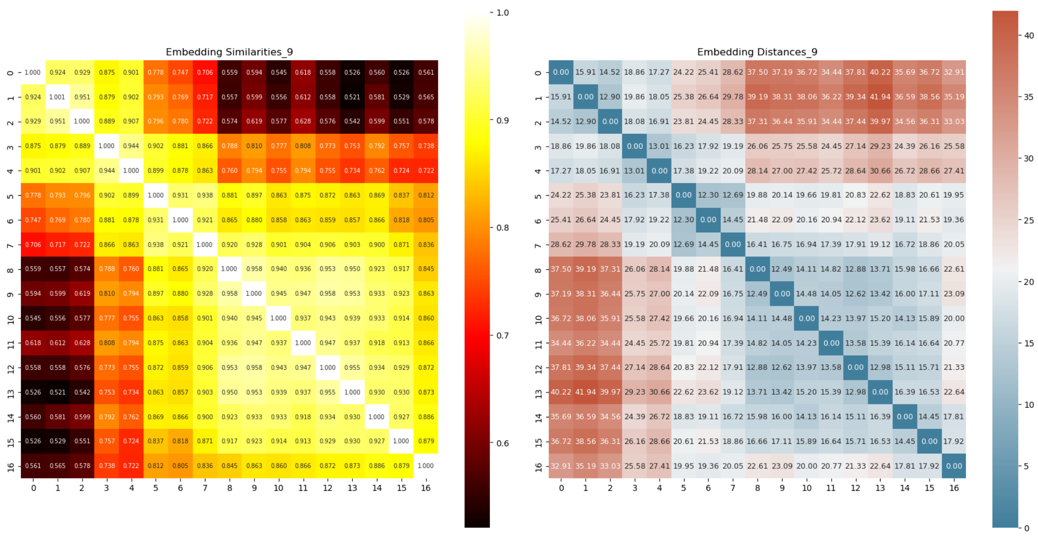

Figure 14: 2.8b Model, Step 9

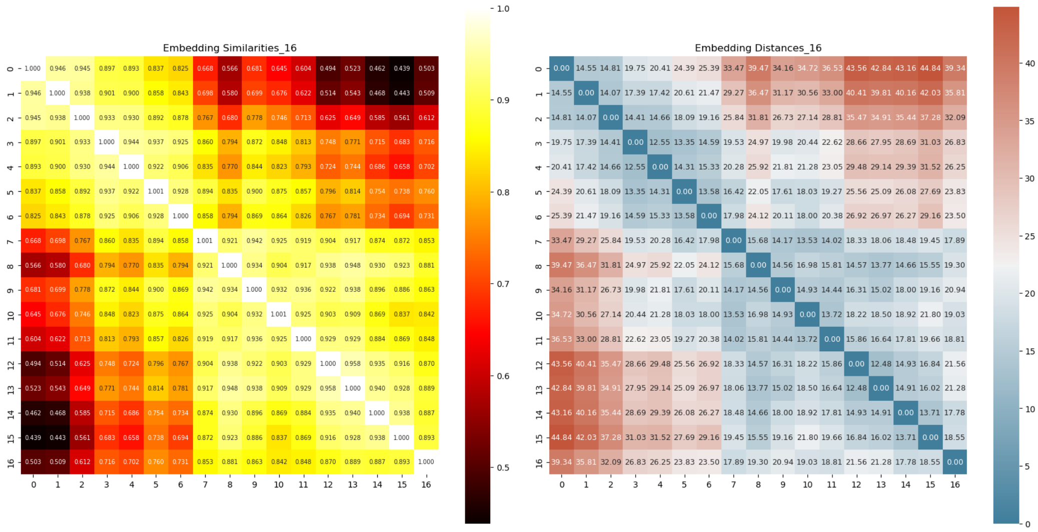

Figure 15: 2.8b Model, Step 16

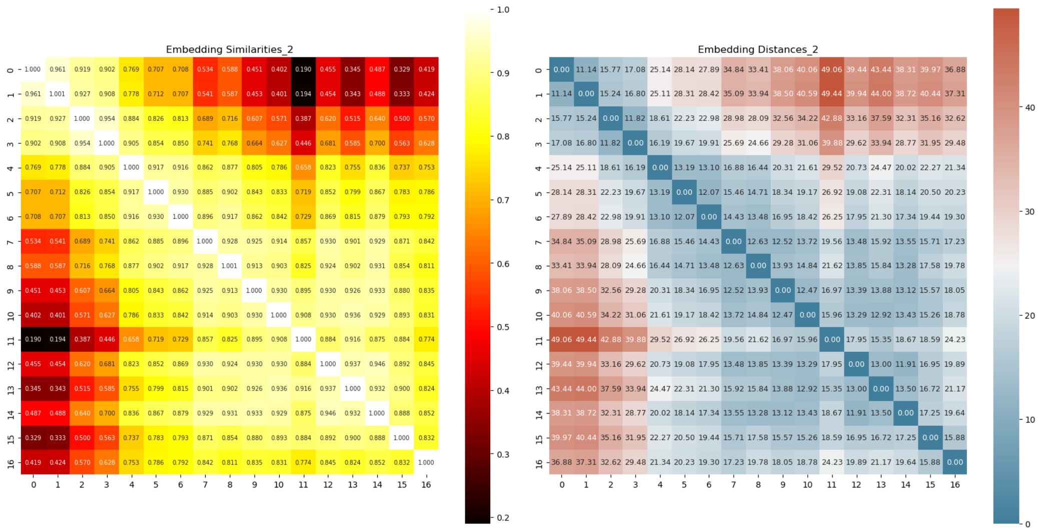

Figure 16: 12b Model, Step 2

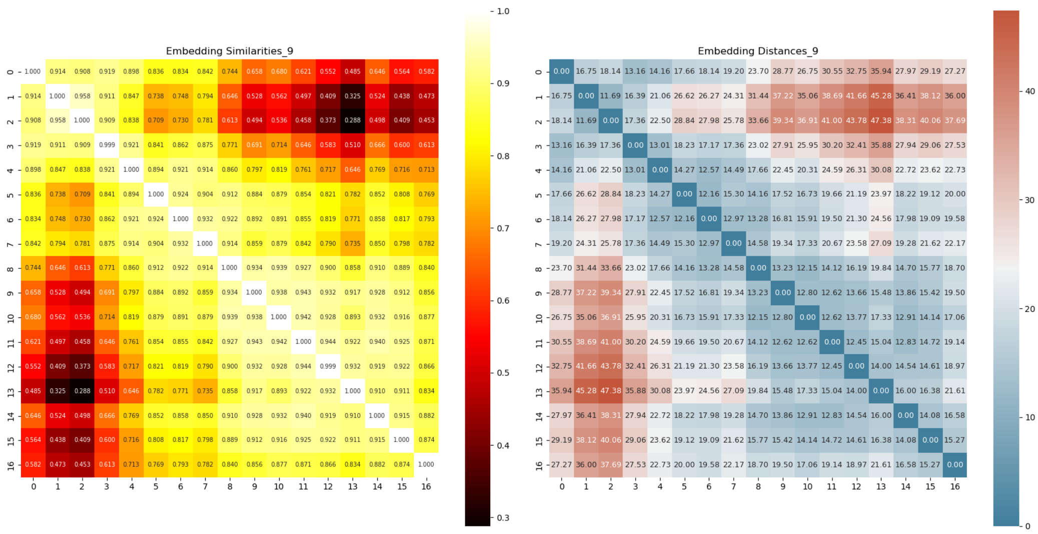

Figure 17: 12b Model, Step 9

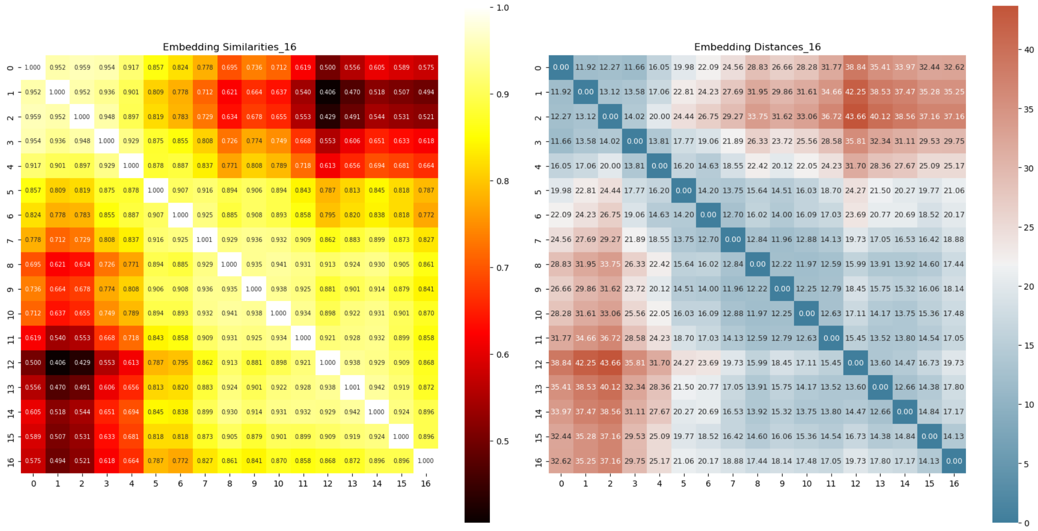

Figure 18: 12b Model, Step 16

not exactly the same but close regarding the memorization score are very close in the embedding space. For example, fully memorized sentences are also close to sentences with high memorization scores and embedding similarity of over 0.9. This shows the possibility that the model is generating paraphrased memorized sequences. However, with the increase in the model size, the cosine similarity decreases. For the Euclidean distance, we can see that the embedding distances have a decreasing trend with the increasing decoding steps, while the mutual Euclidean increases with the model size.

A.4 Does LLM prefer to memorize specific parts within the training data?

In this section, we discuss the question of whether LLM prefers to memorize a specific part within the training data. We split the corpora into 50 parts based on their index and examine how many memorized sentences are in those parts.

From the result shown in Figure 19, though the number of memorized sentences is not completely evenly distributed, we can see that there is no significant part that the number of memorized sentences is clearly more than others. This shows the training order does not affect memorization.

A.5 Detail N-Gram Statistics

We also provided detailed uni-gram and bi-gram statistics for the memorized, un-memorized, half-memorized, and quarter-memorized contents. The average frequency is calculated at each step, and we compute the average frequency of the context, the average frequency of the continuation, and the average frequency of the whole sequence. The results span from the 70m size model to the 12b size model shown in Table 3 to Table 6.

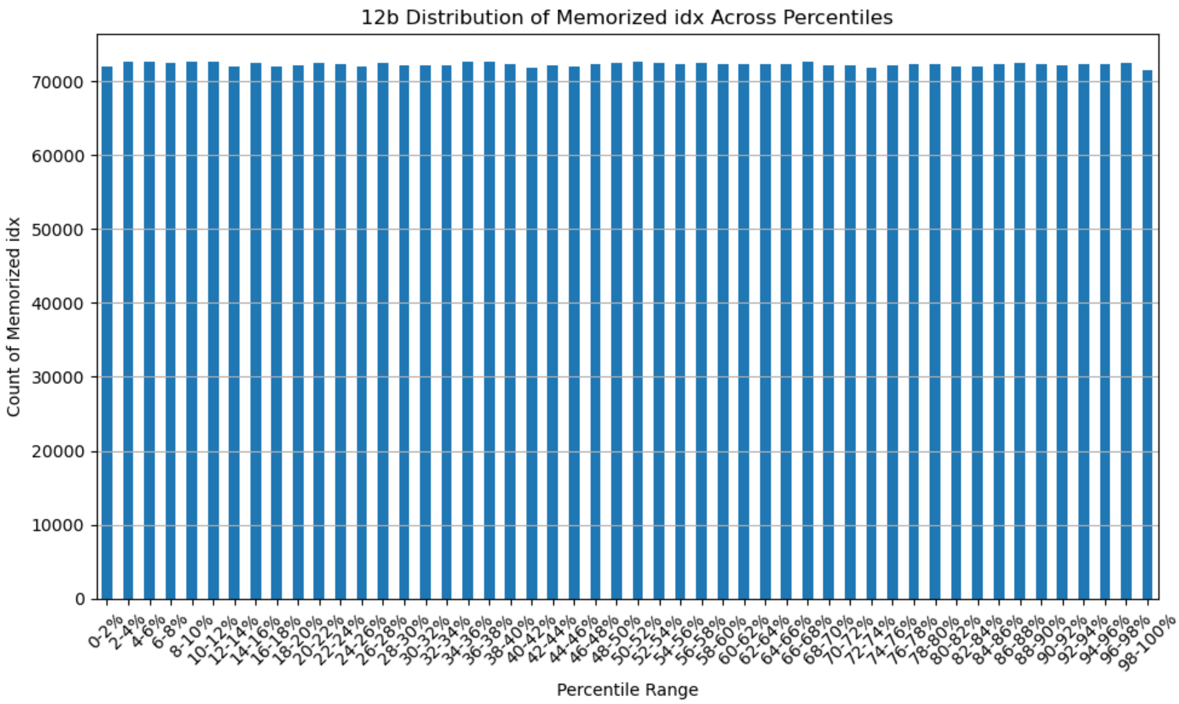

Figure 19: Index Distribution of Memorized Sequences in 12b model.

target size	memorized 70m	memorized 160m	memorized 410m	memorized 1b	memorized 2.8b	memorized 6.9b	memorized 12b	forgotten 70m	forgotten 160m	forgotten 410m	forgotten 1b	forgotten 2.8b	forgotten 6.9b	forgotten 12b
0	1,761,877,325	1,744,677,836	1,753,317,716	1,749,565,941	1,759,384,624	1,763,400,000	1,794,900,000	1,732,531,426	1,732,526,814	1,733,883,954	1,735,430,248	1,757,800,000	1,717,300,000	1,753,200,000
1	1,756,407,896	1,733,466,452	1,737,641,504	1,739,836,645	1,747,576,577	1,712,000,000	1,742,500,000	1,733,790,778	1,734,398,485	1,735,811,740	1,737,110,533	1,754,500,000	1,724,000,000	1,759,900,000
2	1,754,544,442	1,735,828,900	1,739,272,135	1,740,794,079	1,750,863,001	1,758,000,000	1,804,000,000	1,733,955,892	1,734,433,227	1,735,927,786	1,737,263,808	1,719,800,000	1,730,300,000	1,755,600,000
3	1,745,749,714	1,740,145,636	1,739,384,453	1,740,237,878	1,749,816,871	1,700,300,000	1,818,500,000	1,734,815,677	1,735,031,658	1,736,843,088	1,737,987,235	1,686,300,000	1,765,900,000	1,779,800,000
4	1,733,568,727	1,729,484,889	1,728,938,412	1,733,592,149	1,742,113,248	1,766,400,000	1,708,800,000	1,735,128,356	1,735,452,086	1,737,120,799	1,738,793,702	1,767,300,000	1,733,300,000	1,759,200,000
5	1,727,964,181	1,723,277,679	1,725,427,437	1,727,321,616	1,736,693,521	1,738,500,000	1,710,400,000	1,736,270,145	1,736,331,860	1,737,850,119	1,740,340,490	1,758,200,000	1,717,900,000	1,739,900,000
6	1,725,841,351	1,716,563,761	1,723,081,626	1,725,702,469	1,735,415,386	1,758,000,000	1,742,900,000	1,735,559,708	1,736,435,200	1,738,820,970	1,740,165,047	1,780,700,000	1,742,500,000	1,713,400,000
7	1,722,343,456	1,711,523,969	1,719,492,821	1,724,003,560	1,732,499,267	1,698,400,000	1,731,400,000	1,736,139,862	1,737,411,285	1,738,841,087	1,740,979,176	1,708,700,000	1,726,500,000	1,761,400,000
8	1,715,952,971	1,703,541,696	1,712,534,695	1,717,230,465	1,730,417,914	1,679,100,000	1,709,000,000	1,736,938,342	1,738,000,206	1,739,984,548	1,741,715,014	1,752,400,000	1,775,600,000	1,740,500,000
9	1,701,602,676	1,696,501,612	1,705,597,804	1,710,036,650	1,721,980,996	1,776,400,000	1,716,900,000	1,737,218,230	1,738,548,131	1,740,764,383	1,742,163,420	1,759,400,000	1,722,800,000	1,764,800,000
10	1,708,709,485	1,703,507,469	1,707,974,902	1,712,561,959	1,724,060,121	1,672,300,000	1,733,400,000	1,737,059,426	1,738,812,313	1,740,973,658	1,742,911,460	1,760,200,000	1,807,900,000	1,753,800,000
11	1,708,002,048	1,700,334,479	1,705,745,789	1,712,145,483	1,721,887,809	1,719,300,000	1,729,500,000	1,737,667,494	1,739,075,343	1,741,033,716	1,742,470,266	1,743,600,000	1,812,600,000	1,758,700,000
12	1,704,765,267	1,690,528,200	1,700,680,032	1,703,727,332	1,714,582,320	1,739,700,000	1,711,800,000	1,737,833,800	1,738,990,635	1,741,004,691	1,742,710,814	1,768,300,000	1,754,700,000	1,771,800,000
13	1,693,927,960	1,683,230,446	1,695,709,392	1,702,676,948	1,712,420,697	1,731,000,000	1,690,500,000	1,737,172,555	1,738,180,555	1,740,335,149	1,741,471,936	1,698,100,000	1,745,800,000	1,745,100,000
14	1,696,634,570	1,681,852,317	1,695,295,636	1,701,256,215	1,712,257,191	1,725,700,000	1,699,700,000	1,737,522,088	1,738,958,841	1,740,701,751	1,742,059,832	1,740,500,000	1,699,000,000	1,728,500,000
15	1,679,551,926	1,671,142,177	1,688,416,481	1,697,105,648	1,709,669,411	1,752,700,000	1,720,700,000	1,736,261,928	1,737,987,431	1,740,131,961	1,741,684,474	1,752,300,000	1,705,000,000	1,746,200,000
16	1,667,903,751	1,670,413,255	1,678,747,207	1,691,187,011	1,704,670,219	1,719,900,000	1,699,800,000	1,735,533,670	1,737,210,612	1,738,579,355	1,740,344,603	1,750,300,000	1,781,300,000	1,752,600,000
17	1,682,336,989	1,679,636,457	1,684,708,172	1,696,051,548	1,706,414,493	1,718,300,000	1,727,400,000	1,736,647,826	1,737,903,626	1,739,991,371	1,741,869,751	1,754,200,000	1,752,800,000	1,780,200,000
18	1,696,024,732	1,688,157,613	1,693,550,892	1,701,124,435	1,711,161,137	1,689,300,000	1,685,300,000	1,736,875,723	1,738,169,859	1,740,186,111	1,742,150,532	1,723,700,000	1,764,500,000	1,720,000,000
19	1,700,044,430	1,691,379,251	1,700,787,282	1,703,484,170	1,713,713,338	1,735,600,000	1,698,900,000	1,736,115,268	1,738,516,054	1,740,906,141	1,742,896,568	1,716,700,000	1,778,500,000	1,739,500,000
20	1,710,839,951	1,700,701,910	1,706,890,340	1,710,112,768	1,717,662,092	1,703,300,000	1,750,400,000	1,735,637,929	1,737,519,099	1,740,228,710	1,741,971,117	1,776,900,000	1,761,300,000	1,752,300,000
21	1,710,751,717	1,696,906,569	1,701,929,069	1,706,919,694	1,714,832,769	1,726,500,000	1,722,600,000	1,736,494,732	1,739,238,409	1,742,035,789	1,744,490,773	1,747,300,000	1,730,200,000	1,727,900,000
22	1,707,031,048	1,694,596,312	1,700,606,809	1,707,393,546	1,714,688,122	1,688,200,000	1,719,400,000	1,735,343,187	1,738,714,382	1,741,858,543	1,743,975,469	1,763,800,000	1,738,800,000	1,757,900,000
23	1,720,840,503	1,710,023,803	1,709,653,995	1,713,995,947	1,719,166,026	1,728,600,000	1,726,500,000	1,735,401,486	1,738,931,803	1,743,158,826	1,745,299,877	1,733,700,000	1,745,300,000	1,764,500,000
24	1,731,576,174	1,715,940,520	1,714,003,336	1,715,023,292	1,717,926,947	1,758,400,000	1,722,100,000	1,734,596,940	1,738,386,519	1,742,505,032	1,744,500,841	1,749,300,000	1,761,700,000	1,760,400,000
25	1,752,003,292	1,731,855,788	1,727,114,867	1,724,798,309	1,728,919,793	1,724,100,000	1,708,500,000	1,737,772,973	1,742,462,185	1,746,923,721	1,749,250,474	1,773,200,000	1,764,900,000	1,739,600,000
26	1,755,434,381	1,735,918,727	1,728,805,371	1,726,260,802	1,725,855,889	1,717,700,000	1,710,400,000	1,736,254,689	1,741,221,711	1,746,794,790	1,750,257,624	1,782,600,000	1,737,100,000	1,737,500,000
27	1,751,078,472	1,732,474,044	1,721,053,568	1,718,911,936	1,720,139,844	1,730,200,000	1,697,500,000	1,737,465,109	1,745,018,430	1,752,091,624	1,756,426,168	1,732,400,000	1,741,600,000	1,753,500,000
28	1,752,459,582	1,738,590,437	1,733,534,889	1,725,443,806	1,726,659,853	1,686,300,000	1,679,200,000	1,727,033,252	1,738,018,143	1,747,326,896	1,754,100,491	1,764,800,000	1,752,900,000	1,767,200,000
29	1,727,987,920	1,710,624,126	1,706,925,069	1,698,887,785	1,701,059,926	1,699,000,000	1,699,700,000	1,742,590,726	1,758,725,331	1,772,881,062	1,782,612,036	1,789,600,000	1,790,700,000	1,810,600,000
30	1,723,009,291	1,712,857,933	1,705,443,834	1,699,764,161	1,701,619,618	1,711,900,000	1,710,500,000	1,742,138,296	1,764,672,256	1,781,900,380	1,793,731,882	1,863,100,000	1,825,200,000	1,818,400,000
31	1,697,576,775	1,687,242,321	1,674,099,012	1,668,454,050	1,668,926,625	1,649,200,000	1,683,400,000	1,856,009,443	1,902,117,823	1,941,790,973	1,967,891,756	1,975,400,000	1,987,200,000	2,005,800,000
Avg Context	1,719,510,719	1,708,216,456	1,711,448,892	1,713,925,259	1,721,720,489	1,721,178,125	1,722,078,125	1,739,930,530	1,744,606,260	1,749,349,648	1,752,719,607	1,759,534,375	1,759,221,875	1,763,115,625
32	1,812,724,891	1,801,299,369	1,783,560,981	1,771,127,395	1,765,849,310	1,739,200,000	1,722,000,000	957,760,109	963,964,253	981,205,278	986,644,973	1,017,100,000	1,024,300,000	989,330,000
33	1,785,198,082	1,782,599,042	1,764,708,279	1,754,286,257	1,750,603,678	1,748,500,000	1,734,000,000	1,477,089,688	1,452,367,316	1,435,934,090	1,428,928,547	1,415,800,000	1,418,800,000	1,417,900,000
34	1,772,561,111	1,767,491,083	1,753,649,952	1,744,671,079	1,742,356,129	1,769,300,000	1,735,800,000	1,623,949,166	1,609,532,509	1,597,061,854	1,592,797,469	1,560,600,000	1,576,000,000	1,592,600,000
35	1,762,468,180	1,755,171,008	1,742,357,780	1,734,034,726	1,733,128,170	1,732,800,000	1,735,800,000	1,656,747,957	1,650,264,891	1,644,001,022	1,641,698,771	1,625,200,000	1,654,900,000	1,635,900,000
36	1,751,084,744	1,744,819,021	1,731,757,083	1,727,988,589	1,726,739,343	1,728,600,000	1,706,600,000	1,678,449,622	1,673,912,651	1,669,885,767	1,668,314,472	1,642,100,000	1,683,100,000	1,643,100,000
37	1,737,469,976	1,729,200,256	1,720,955,403	1,720,387,855	1,719,147,209	1,712,900,000	1,787,300,000	1,694,329,815	1,691,870,813	1,689,952,951	1,689,876,998	1,727,300,000	1,699,700,000	1,699,200,000
38	1,737,353,894	1,724,369,474	1,716,962,208	1,710,620,923	1,711,850,544	1,691,900,000	1,761,400,000	1,696,281,632	1,695,900,348	1,694,668,623	1,695,468,826	1,674,600,000	1,697,300,000	1,679,200,000
39	1,715,081,363	1,703,843,951	1,703,301,172	1,705,983,119	1,709,368,578	1,706,400,000	1,660,200,000	1,700,949,641	1,700,918,018	1,700,662,603	1,701,381,229	1,735,200,000	1,703,000,000	1,694,800,000
40	1,686,685,834	1,682,653,925	1,690,744,526	1,694,008,976	1,698,953,947	1,698,400,000	1,732,100,000	1,699,950,424	1,700,320,136	1,700,698,424	1,701,888,972	1,709,600,000	1,749,200,000	1,717,200,000
41	1,697,887,078	1,690,596,776	1,699,039,050	1,702,293,288	1,706,226,135	1,706,000,000	1,729,200,000	1,701,980,579	1,703,673,165	1,704,305,693	1,705,426,679	1,713,800,000	1,688,200,000	1,721,300,000
42	1,706,446,759	1,699,625,915	1,704,257,959	1,710,138,829	1,711,973,193	1,706,800,000	1,696,400,000	1,699,916,580	1,701,963,362	1,704,179,446	1,704,239,602	1,693,500,000	1,680,700,000	1,699,700,000
43	1,724,996,796	1,711,048,851	1,718,405,094	1,721,177,377	1,728,269,651	1,730,100,000	1,737,300,000	1,701,424,868	1,702,896,473	1,704,910,899	1,706,115,083	1,699,300,000	1,718,400,000	1,702,000,000
44	1,729,978,559	1,717,877,194	1,721,748,175	1,727,317,540	1,725,798,564	1,752,400,000	1,680,700,000	1,699,235,317	1,702,231,574	1,704,286,186	1,705,328,377	1,732,700,000	1,719,100,000	1,715,900,000
45	1,756,734,060	1,740,464,386	1,745,501,552	1,750,234,915	1,752,814,925	1,722,000,000	1,740,000,000	1,699,769,885	1,702,350,463	1,703,950,116	1,705,365,109	1,683,000,000	1,753,700,000	1,696,100,000
46	1,772,357,723	1,765,976,491	1,769,251,508	1,775,699,434	1,780,492,152	1,796,400,000	1,773,900,000	1,697,965,233	1,700,236,541	1,702,146,238	1,703,508,158	1,740,600,000	1,727,900,000	1,724,700,000
47	1,826,784,956	1,810,274,933	1,823,557,511	1,829,605,754	1,834,611,860	1,849,800,000	1,853,200,000	1,694,816,806	1,697,715,045	1,699,839,943	1,701,149,553	1,730,400,000	1,716,300,000	1,675,600,000
Avg Continuation	1,748,488,375	1,739,206,980	1,736,859,890	1,736,223,503	1,737,386,462	1,736,968,750	1,736,618,750	1,630,038,582	1,628,132,347	1,627,355,571	1,627,383,301	1,631,300,000	1,638,162,500	1,625,283,125
Avg Sentence	1,729,169,938	1,723,124,945	1,719,919,225	1,721,358,007	1,726,942,480	1,726,441,667	1,726,925,000	1,703,299,881	1,705,781,622	1,708,684,955	1,710,940,838	1,716,789,583	1,718,868,750	1,717,171,458

Table 3: Uni-gram Statistics for Memorized and Unmemorized Content

target size	half 70m	half 160m	half 410m	half 1b	half 2.8b	half 6.9b	half 12b	quarter 70m	quarter 160m	quarter 410m	quarter 1b	quarter 2.8b	quarter 6.9b	quarter 12b
0	1,747,500,000	1,780,200,000	1,752,900,000	1,700,100,000	1,729,500,000	1,733,700,000	1,715,500,000	1774800000	1,754,800,000	1785800000	1763800000	1781400000	1778900000	1756200000
1	1,771,400,000	1,735,900,000	1,707,600,000	1,717,800,000	1,756,600,000	1,731,400,000	1,729,600,000	1746500000	1,781,100,000	1767400000	1746500000	1742600000	1779500000	1736400000
2	1,747,800,000	1,745,400,000	1,758,500,000	1,731,800,000	1,705,000,000	1,687,200,000	1,774,100,000	1754000000	1,794,900,000	1781500000	1756500000	1768900000	1760300000	1753800000
3	1,760,400,000	1,712,100,000	1,743,700,000	1,745,200,000	1,726,600,000	1,753,500,000	1,763,800,000	1791200000	1,776,600,000	1769500000	1776000000	1791900000	1748200000	1782600000
4	1,693,800,000	1,720,700,000	1,726,500,000	1,696,900,000	1,748,200,000	1,697,300,000	1,723,500,000	1696800000	1,740,300,000	1789400000	1761900000	1792900000	1760800000	1754400000
5	1,711,200,000	1,746,400,000	1,762,300,000	1,722,900,000	1,680,700,000	1,715,700,000	1,714,000,000	1761700000	1,765,300,000	1777000000	1720700000	1763800000	1765800000	1746100000
6	1,700,900,000	1,717,100,000	1,673,600,000	1,753,500,000	1,774,500,000	1,735,800,000	1,771,000,000	1764900000	1,757,900,000	1726200000	1744400000	1742400000	1761400000	1785200000
7	1,737,400,000	1,743,900,000	1,710,200,000	1,741,300,000	1,766,200,000	1,688,400,000	1,730,600,000	1755400000	1,690,800,000	1790000000	1717000000	1751600000	1740100000	1754900000
8	1,705,800,000	1,721,300,000	1,728,200,000	1,723,200,000	1,720,900,000	1,676,100,000	1,738,200,000	1739900000	1,726,100,000	1799800000	1773900000	1806700000	1704900000	1737600000
9	1,754,600,000	1,756,100,000	1,762,300,000	1,722,400,000	1,767,400,000	1,772,800,000	1,708,800,000	1785800000	1,741,600,000	1756200000	1763000000	1744900000	1757800000	1708900000
10	1,737,700,000	1,726,600,000	1,761,000,000	1,715,500,000	1,724,600,000	1,754,200,000	1,757,500,000	1753600000	1,763,700,000	1757000000	1760600000	1748400000	1736600000	1770700000
11	1,778,500,000	1,738,600,000	1,691,800,000	1,753,500,000	1,704,100,000	1,744,300,000	1,765,900,000	1774300000	1,773,600,000	1764500000	1746700000	1752300000	1741700000	1740200000
12	1,721,500,000	1,696,300,000	1,678,400,000	1,699,000,000	1,714,100,000	1,718,900,000	1,765,400,000	1776700000	1,734,000,000	1771900000	1716500000	1736100000	1727900000	1730800000
13	1,751,300,000	1,732,400,000	1,708,000,000	1,700,400,000	1,698,800,000	1,754,600,000	1,719,200,000	1714300000	1,744,900,000	1749500000	1741100000	1729200000	1746400000	1742400000
14	1,737,600,000	1,731,900,000	1,735,500,000	1,682,700,000	1,695,200,000	1,701,600,000	1,730,800,000	1751000000	1,749,400,000	1731600000	1773600000	1738200000	1766800000	1712100000
15	1,693,700,000	1,725,700,000	1,774,900,000	1,732,300,000	1,713,300,000	1,775,300,000	1,743,900,000	1728000000	1,724,300,000	1739000000	1742500000	1762500000	1736200000	1782100000
16	1,739,000,000	1,749,900,000	1,740,800,000	1,777,700,000	1,747,300,000	1,749,300,000	1,702,800,000	1738500000	1,740,800,000	1705500000	1771500000	1729500000	1763500000	1749200000
17	1,741,900,000	1,694,200,000	1,742,100,000	1,771,200,000	1,721,200,000	1,741,300,000	1,783,000,000	1745700000	1,740,100,000	1767500000	1736700000	1738900000	1741000000	1761400000
18	1,738,200,000	1,703,300,000	1,740,700,000	1,711,900,000	1,705,700,000	1,704,600,000	1,686,400,000	1737100000	1,743,400,000	1743200000	1737100000	1724500000	1735100000	1756800000
19	1,728,500,000	1,725,600,000	1,729,100,000	1,719,600,000	1,725,700,000	1,679,600,000	1,747,000,000	1726900000	1,791,300,000	1728300000	1740300000	1707300000	1736600000	1719900000
20	1,691,300,000	1,761,400,000	1,714,800,000	1,761,200,000	1,695,900,000	1,702,800,000	1,733,400,000	1727700000	1,767,400,000	1707000000	1730800000	1771400000	1725200000	1795800000
21	1,677,600,000	1,738,600,000	1,699,400,000	1,737,100,000	1,691,400,000	1,680,900,000	1,697,700,000	1755200000	1,741,500,000	1750000000	1777700000	1751000000	1741700000	1746900000
22	1,740,300,000	1,713,600,000	1,716,200,000	1,668,000,000	1,698,100,000	1,718,800,000	1,722,600,000	1774400000	1,710,600,000	1757000000	1774700000	1797700000	1752900000	1746100000
23	1,748,200,000	1,666,100,000	1,664,400,000	1,729,100,000	1,677,100,000	1,734,500,000	1,731,200,000	1747300000	1,737,000,000	1766600000	1782700000	1761900000	1757800000	1723400000
24	1,691,700,000	1,735,300,000	1,716,500,000	1,722,300,000	1,733,100,000	1,699,600,000	1,678,600,000	1735800000	1,734,200,000	1748600000	1712500000	1743000000	1772400000	1700900000
25	1,728,800,000	1,656,600,000	1,726,600,000	1,700,200,000	1,690,500,000	1,697,900,000	1,732,800,000	1731200000	1,766,800,000	1740300000	1724100000	1736800000	1782600000	1718200000
26	1,700,700,000	1,715,200,000	1,654,200,000	1,694,200,000	1,703,700,000	1,684,400,000	1,733,500,000	1708800000	1,715,100,000	1739000000	1734800000	1721100000	1695100000	1749000000
27	1,724,500,000	1,645,400,000	1,754,400,000	1,707,800,000	1,724,900,000	1,721,800,000	1,672,900,000	1703300000	1,715,900,000	1711100000	1697200000	1718800000	1708100000	1695900000
28	1,658,300,000	1,712,500,000	1,744,100,000	1,660,800,000	1,739,500,000	1,678,300,000	1,701,000,000	1743400000	1,732,800,000	1752700000	1712300000	1724300000	1736400000	1680500000
29	1,692,600,000	1,667,500,000	1,686,700,000	1,655,700,000	1,657,100,000	1,676,700,000	1,707,500,000	1697300000	1,771,300,000	1731300000	1717800000	1677500000	1714100000	1691500000
30	1,655,500,000	1,690,700,000	1,658,200,000	1,647,300,000	1,683,200,000	1,639,600,000	1,628,200,000	1760800000	1,689,800,000	1675800000	1691500000	1678700000	1649800000	1697800000
31	1,666,800,000	1,541,100,000	1,637,100,000	1,564,000,000	1,513,400,000	1,579,400,000	1,554,100,000	1562100000	1,515,300,000	1536200000	1484100000	1477500000	1436800000	1507600000
Avg Context	1,721,093,750	1,713,987,500	1,718,771,875	1,711,456,250	1,710,421,875	1,710,321,875	1,720,765,625	1,739,512,500	1,738,518,750	1,744,262,500	1,735,328,125	1,737,928,125	1,733,200,000	1,732,353,125
32	1,878,300,000	1,871,300,000	1,800,300,000	1,834,700,000	1,812,500,000	1,719,500,000	1,791,700,000	2189100000	2,090,900,000	2020500000	1992900000	1907700000	1938700000	1895000000
33	1,842,000,000	1,854,700,000	1,836,200,000	1,829,200,000	1,869,400,000	1,834,500,000	1,749,400,000	2458900000	2,344,000,000	2383700000	2295800000	2224900000	2269700000	2275200000
34	1,842,600,000	1,900,800,000	1,843,500,000	1,862,600,000	1,778,400,000	1,812,100,000	1,832,600,000	2650800000	2,597,500,000	2592700000	2537800000	2483300000	2511000000	2568800000
35	1,841,600,000	1,875,500,000	1,883,700,000	1,894,300,000	1,866,800,000	1,843,400,000	1,868,600,000	2520900000	2,583,300,000	2586100000	2591600000	2586700000	2599000000	2572600000
36	1,871,800,000	1,901,500,000	1,932,600,000	1,931,300,000	1,926,100,000	1,872,300,000	1,893,200,000	1778900000	1,742,400,000	1743700000	1696700000	1687300000	1715400000	1720500000
37	1,885,600,000	1,921,800,000	1,997,900,000	1,915,800,000	1,981,000,000	2,077,100,000	2,038,100,000	1736300000	1,739,900,000	1728100000	1685100000	1664500000	1681500000	1700300000
38	1,998,200,000	2,101,300,000	2,081,700,000	2,115,900,000	2,156,700,000	2,209,100,000	2,254,500,000	1762300000	1,710,100,000	1711100000	1737600000	1718500000	1736000000	1705300000
39	2,032,100,000	2,158,400,000	2,168,100,000	2,189,200,000	2,246,300,000	2,276,200,000	2,369,000,000	1809200000	1,768,100,000	1799700000	1747600000	1744100000	1738300000	1758800000
40	1,723,200,000	1,690,600,000	1,715,300,000	1,692,500,000	1,691,600,000	1,691,900,000	1,719,100,000	1765200000	1,751,200,000	1789500000	1746300000	1757000000	1733900000	1779400000
41	1,738,700,000	1,731,600,000	1,699,700,000	1,709,900,000	1,706,600,000	1,668,700,000	1,706,000,000	1773400000	1,777,100,000	1770200000	1803300000	1780300000	1788700000	1780500000
42	1,707,400,000	1,795,200,000	1,719,700,000	1,721,300,000	1,743,300,000	1,700,900,000	1,725,800,000	1822600000	1,814,600,000	1763300000	1800800000	1793500000	1788700000	1740000000
43	1751700000	1,708,000,000	1,720,500,000	1,713,600,000	1,733,400,000	1,727,600,000	1,687,200,000	1839800000	1,774,600,000	1838300000	1761600000	1754700000	1787300000	1805000000
44	1753600000	1,742,000,000	1,764,800,000	1,694,000,000	1,739,700,000	1,686,500,000	1,721,000,000	1760300000	1,806,500,000	1817700000	1760600000	1787800000	1782700000	1752900000
45	1729900000	1,720,800,000	1,715,200,000	1,721,700,000	1,786,500,000	1,697,200,000	1,733,700,000	1775700000	1,763,900,000	1829800000	1840800000	1784300000	1822300000	1734000000
46	1733000000	1,717,200,000	1,741,500,000	1,705,100,000	1,750,600,000	1,721,000,000	1,737,400,000	1797500000	1,799,100,000	1788600000	1729000000	1772000000	1765100000	1758700000
47	1732500000	1,715,600,000	1,748,500,000	1,788,200,000	1,710,100,000	1,731,500,000	1,712,500,000	1763400000	1,757,300,000	1783200000	1835300000	1791700000	1764800000	1766000000
Avg Continuation	1,816,387,500	1,837,893,750	1,835,575,000	1,832,456,250	1,843,687,500	1,829,343,750	1,846,237,500	1,950,268,750	1,926,281,250	1,934,137,500	1,910,175,000	1,889,893,750	1,901,443,750	1,894,562,500
Avg Sentence	1,752,858,333	1,755,289,583	1,757,706,250	1,751,789,583	1,754,843,750	1,749,995,833	1,762,589,583	1,809,764,583	1,801,106,250	1,807,554,167	1,793,610,417	1,788,583,333	1,789,281,250	1,786,422,917

Table 4: Uni-gram Statistics for Half-memorized and Quarter-memorized content

target size	memorized 70m	memorized 160m	memorized 410m	memorized 1b	memorized 2.8b	memorized 6.9b	memorized 12b	forgotten 70m	forgotten 160m	forgotten 410m	forgotten 1b	forgotten 2.8b	forgotten 6.9b	forgotten 12b
1	56,865,298	55,633,408	57,631,394	57,851,391	58,747,419	59,390,492	59,538,566	67,723,858	67,753,352	67,683,530	67,708,452	67,740,430	67,754,040	67,767,823
2	57,323,714	56,317,639	57,433,612	57,993,010	58,602,181	59,488,016	59,803,389	67,737,009	67,768,452	67,769,364	67,815,489	67,769,957	67,800,825	67,828,380
3	56,730,613	56,231,828	57,388,365	58,009,821	58,928,771	59,525,329	59,598,379	67,775,953	67,738,481	67,807,016	67,780,588	67,804,139	67,851,921	67,894,036
4	55,911,242	56,234,359	56,980,241	57,792,288	58,806,766	59,391,815	59,636,502	67,949,220	67,825,207	67,839,517	67,841,697	67,867,488	67,909,524	67,967,557
5	55,692,756	55,655,983	56,819,836	57,273,741	58,351,556	59,022,366	59,346,451	67,958,033	67,886,128	67,874,628	68,003,000	67,965,687	68,045,776	68,030,203
6	55,894,920	55,666,111	56,911,193	57,438,723	58,686,676	59,320,405	59,557,076	67,985,473	67,869,913	67,917,993	68,012,109	67,974,605	68,003,875	68,007,874
7	56,047,068	55,753,285	56,733,219	57,360,465	58,300,186	59,087,792	59,572,316	68,058,487	68,015,919	68,064,231	68,101,370	68,115,977	68,153,223	68,184,607
8	55,832,397	55,349,442	56,329,311	57,111,974	58,349,971	58,924,645	59,505,541	68,017,225	68,013,215	68,034,561	68,111,358	68,101,007	68,139,029	68,201,839
9	55,253,643	54,713,056	55,885,459	56,532,115	57,837,312	58,610,331	59,086,863	68,135,592	68,114,172	68,143,152	68,214,493	68,138,035	68,189,743	68,155,354
10	54,503,716	54,257,056	55,495,149	56,166,844	57,307,569	58,022,654	58,377,791	68,160,293	68,172,545	68,187,816	68,250,398	68,215,884	68,267,972	68,242,452
11	54,988,987	54,708,110	55,773,288	56,333,064	57,701,866	58,500,535	58,871,367	68,234,320	68,217,765	68,269,026	68,248,385	68,243,497	68,268,692	68,241,732
12	54,935,679	54,664,019	55,533,124	56,267,162	57,079,217	58,003,846	58,331,501	68,213,578	68,234,560	68,252,214	68,247,968	68,328,780	68,343,161	68,299,170
13	54,767,327	54,145,183	55,296,145	55,857,794	56,871,961	57,652,393	57,921,174	68,226,649	68,274,620	68,277,162	68,277,302	68,337,296	68,382,126	68,389,974
14	54,021,320	53,569,285	54,624,126	55,455,249	56,409,277	57,387,470	57,829,909	68,213,386	68,233,133	68,245,257	68,296,277	68,307,432	68,429,171	68,395,506
15	54,367,150	53,970,143	55,279,001	55,940,451	57,010,866	57,759,141	58,109,891	68,177,596	68,207,430	68,237,580	68,241,325	68,279,727	68,326,632	68,368,250
16	53,451,225	53,327,809	54,393,699	55,350,104	56,364,900	57,290,723	57,711,596	68,022,941	68,094,829	68,169,848	68,158,831	68,224,797	68,273,585	68,255,991
17	53,645,203	53,634,633	54,177,271	55,078,600	56,138,436	57,251,531	57,541,351	68,138,549	68,147,333	68,221,638	68,287,831	68,325,084	68,415,608	68,411,921
18	54,668,363	54,174,847	55,256,176	56,064,089	56,740,259	57,769,268	57,973,134	68,224,597	68,282,517	68,395,059	68,437,841	68,522,898	68,598,086	68,599,844
19	55,562,101	54,791,623	55,294,695	55,928,589	56,768,136	57,599,180	58,081,287	68,209,538	68,324,444	68,446,514	68,525,618	68,628,926	68,686,765	68,753,261
20	55,938,657	55,081,208	55,877,542	56,253,266	57,048,367	57,967,318	58,106,207	68,136,189	68,313,557	68,470,777	68,543,198	68,639,292	68,722,909	68,829,286
21	56,331,087	55,229,266	55,652,638	56,067,216	56,679,203	57,502,276	57,635,297	68,141,999	68,361,062	68,547,360	68,644,264	68,783,591	68,859,312	68,950,347
22	56,997,804	55,715,430	56,053,895	56,356,954	56,761,453	57,545,255	57,751,704	68,198,529	68,480,921	68,670,594	68,818,205	69,044,454	69,163,424	69,194,391
23	56,802,977	55,429,313	55,735,965	56,197,410	56,471,773	57,057,789	57,383,903	68,143,528	68,525,019	68,831,257	68,946,803	69,230,759	69,321,426	69,445,410
24	57,826,178	56,183,113	56,044,365	55,945,228	55,984,317	56,618,627	56,771,473	68,271,377	68,728,185	69,037,189	69,197,403	69,488,552	69,616,627	69,731,068
25	58,617,239	57,079,668	56,202,206	56,102,928	55,885,915	56,296,069	56,349,525	68,384,381	68,871,654	69,255,113	69,446,599	69,840,923	69,999,312	70,097,498
26	59,309,065	57,207,648	56,075,430	55,816,944	55,552,723	55,814,962	55,846,995	68,608,814	69,166,616	69,653,434	69,885,998	70,350,797	70,565,524	70,654,203
27	59,214,372	56,581,831	55,453,438	55,266,849	54,772,970	55,091,733	55,084,942	68,827,806	69,591,707	70,264,503	70,708,424	71,312,980	71,558,722	71,777,086
28	58,905,505	56,371,887	55,174,878	54,389,172	54,041,511	54,335,228	54,268,336	68,505,025	69,619,124	70,499,723	71,115,782	71,913,592	72,172,923	72,454,327
29	58,023,696	56,122,492	55,053,358	53,998,472	53,301,028	53,274,003	53,087,300	68,732,220	70,492,616	71,644,050	72,524,388	73,553,428	73,967,910	74,302,156
30	56,113,860	54,313,900	52,636,358	51,531,025	50,770,809	50,519,079	50,314,047	74,810,707	77,539,372	79,590,655	80,895,889	82,592,009	83,417,356	83,993,892
31	51,887,230	50,597,329	48,600,945	47,771,643	46,600,073	46,175,796	45,915,086	77,719,313	81,332,604	84,241,907	86,088,401	88,563,161	89,838,828	90,658,855
Avg Context	56,013,884	55,119,707	55,541,817	55,854,922	56,415,273	57,038,583	57,255,126	68,698,135	69,103,111	69,436,860	69,657,280	69,942,103	70,098,195	70,196,268
32	55,823,256	53,884,097	51,681,052	50,569,791	49,394,788	48,863,702	48,537,557	13,930,348	15,076,103	16,016,778	16,465,745	16,788,294	17,279,969	17,455,400
33	58,738,344	56,674,056	54,054,542	52,888,037	51,528,984	50,887,306	50,414,778	26,816,835	27,093,410	27,773,603	27,962,715	28,191,993	28,477,352	28,654,271
34	56,938,764	55,281,907	52,909,883	51,461,738	50,038,303	49,641,208	49,084,752	47,564,034	45,235,994	44,031,407	43,883,655	43,877,411	43,554,253	43,485,497
35	54,366,396	53,029,014	50,888,507	49,755,655	48,770,433	48,334,318	47,981,902	53,101,160	51,120,417	49,742,735	49,185,432	48,518,180	48,081,379	47,894,499
36	52,936,332	51,682,636	49,570,706	48,713,389	47,794,113	47,573,073	47,233,332	58,003,949	56,939,073	55,963,819	55,597,113	55,051,241	54,774,326	54,588,795
37	51,647,223	50,371,630	48,303,608	47,632,078	46,858,461	46,786,053	46,376,421	61,213,159	60,453,961	59,829,451	59,536,354	59,153,093	59,031,495	58,838,218
38	50,931,788	49,209,356	47,576,761	46,731,794	45,968,874	45,790,581	45,477,362	62,655,390	62,137,314	61,620,255	61,458,679	61,231,410	61,110,159	61,003,516
39	51,042,601	49,027,546	47,618,052	46,768,291	45,998,476	45,883,841	45,473,705	63,348,201	62,998,381	62,683,843	62,649,981	62,440,555	62,378,970	62,319,582
40	49,590,360	47,756,219	46,431,515	45,965,206	45,449,164	45,252,564	45,184,300	63,866,857	63,600,670	63,423,587	63,430,422	63,301,531	63,282,617	63,207,398
41	49,954,374	47,928,574	46,848,807	46,364,193	45,679,210	45,474,688	45,440,294	64,301,293	64,110,817	63,913,386	63,893,724	63,873,160	63,880,434	63,857,697
42	50,897,285	48,822,840	47,723,273	47,389,348	46,461,956	46,336,541	46,106,581	64,486,391	64,515,591	64,453,381	64,379,111	64,406,903	64,442,057	64,385,303
43	52,467,447	49,948,494	48,687,532	48,087,823	47,426,244	47,134,027	46,970,169	64,557,953	64,566,576	64,532,662	64,488,626	64,565,552	64,573,767	64,585,706
44	54,876,822	52,145,505	50,513,749	49,816,893	48,819,838	48,428,789	48,281,333	64,662,055	64,739,419	64,747,685	64,807,289	64,861,456	64,982,075	64,897,827
45	56,056,772	53,288,677	51,530,428	50,709,794	49,538,754	49,171,650	48,847,311	64,772,578	64,907,757	64,948,826	64,997,734	65,068,658	65,124,986	65,132,858
46	61,167,369	59,469,476	57,204,341	55,850,490	54,489,360	53,898,637	53,288,932	64,727,227	64,840,307	64,926,333	64,908,159	65,053,638	65,091,642	65,154,625
47	71,009,467	67,839,176	66,565,545	65,503,549	63,999,825	63,292,284	62,803,150	64,590,446	64,867,523	64,840,482	64,903,198	65,014,873	65,036,948	65,120,427
48	66,672,308	64,419,151	63,096,912	62,676,497	61,761,157	60,991,890	60,750,602	66,612,523	66,803,144	66,905,353	66,952,648	67,075,374	67,116,770	67,266,114
Avg Continuation	55,595,112	53,575,197	51,835,601	50,993,210	49,998,702	49,631,832	49,308,969	57,012,377	56,706,262	56,491,387	56,441,211	56,380,784	56,365,835	56,343,984
Avg Sentence	55,865,569	54,572,693	54,229,199	54,133,066	54,142,738	54,415,359	54,440,862	64,559,429	64,712,561	64,852,005	64,976,589	65,139,135	65,234,651	65,290,251

Table 5: Bi-gram Statistics for memorized and unmemorized content

target size	half 70m	half 160m	half 410m	half 1b	half 2.8b	half 6.9b	half 12b	quarter 70m	quarter 160m	quarter 410m	quarter 1b	quarter 2.8b	quarter 6.9b	quarter 12b
1	53,888,945	55,774,256	56,835,849	57,399,791	58,749,349	58,746,839	59,572,656	65,365,004	66,059,083	66,813,871	67,057,786	67,482,982	67,249,886	67,522,493
2	53,633,778	55,585,291	56,294,740	57,125,864	58,375,426	58,787,083	59,024,756	65,505,804	66,197,266	66,585,597	66,774,723	67,277,478	67,415,528	67,313,805
3	53,966,515	55,464,751	56,571,797	57,248,585	58,522,726	59,007,609	59,282,618	65,526,925	66,301,214	66,609,558	67,101,450	67,242,630	67,240,433	67,479,250
4	53,788,979	55,012,249	55,852,524	56,732,428	58,632,577	58,585,481	59,106,526	65,575,121	65,977,326	66,498,055	66,914,635	67,144,082	67,128,936	67,304,810
5	52,887,357	54,130,559	55,380,593	56,487,949	57,789,441	58,410,016	58,836,366	65,413,320	66,086,522	66,640,160	66,821,619	67,103,545	67,257,537	67,368,841
6	53,346,585	54,626,718	56,348,320	56,603,397	58,159,466	58,648,425	58,945,615	65,113,237	66,153,641	66,499,304	66,682,733	67,063,613	67,252,679	67,384,155
7	52,836,947	54,888,957	55,457,422	56,613,020	57,855,215	58,349,009	58,580,014	65,090,106	65,649,597	66,459,689	66,678,897	66,947,213	67,002,958	67,127,891
8	52,478,474	54,201,487	55,628,345	56,114,532	57,493,118	58,018,440	58,920,964	64,824,999	65,498,431	66,241,656	66,399,994	66,746,465	66,760,013	66,819,417
9	53,600,100	55,410,940	56,362,313	57,618,326	58,256,347	59,088,717	59,814,719	64,941,905	65,788,899	66,460,344	66,506,701	66,921,900	67,080,681	67,174,870
10	53,868,058	55,087,199	56,362,786	57,058,982	58,314,177	58,745,840	59,526,404	64,563,380	65,502,285	66,335,635	66,276,258	66,961,511	66,922,326	66,965,502
11	52,884,280	54,383,701	56,075,362	56,818,188	58,050,191	58,352,813	58,596,784	64,715,807	65,616,229	66,029,659	66,452,694	66,712,759	66,674,522	66,995,794
12	52,928,646	54,849,387	55,982,049	56,771,448	58,094,228	58,565,819	58,811,948	64,820,938	65,424,995	65,790,449	66,253,435	66,525,526	66,698,298	66,719,967
13	52,690,921	54,559,756	55,697,194	56,532,692	57,846,992	58,355,202	58,946,804	64,842,590	65,142,936	65,893,506	66,065,102	66,465,922	66,738,926	66,653,765
14	52,781,679	54,368,904	56,019,791	56,847,208	57,869,958	58,242,828	58,685,929	64,404,765	65,245,584	65,797,393	66,052,000	66,479,573	66,541,779	66,586,451
15	53,765,298	55,537,152	56,289,403	57,279,213	58,113,960	58,741,123	59,494,649	64,231,835	64,970,624	65,603,603	65,691,554	66,258,007	66,393,008	66,421,189
16	53,740,078	55,037,415	55,929,818	56,941,982	57,702,245	58,282,107	58,697,163	64,285,467	64,973,050	65,644,470	65,985,554	66,366,943	66,449,639	66,680,453
17	53,610,921	55,130,291	56,781,859	57,648,192	58,161,184	58,670,246	59,044,372	64,584,535	65,324,214	66,141,495	66,173,896	66,423,703	66,634,663	66,523,072
18	52,403,931	54,507,436	56,143,121	56,609,104	57,872,099	58,070,077	58,324,067	64,823,801	65,426,701	65,725,421	65,794,806	66,079,791	66,355,245	66,570,526
19	51,734,926	53,970,664	55,420,133	56,429,139	57,389,684	58,026,044	57,807,255	64,435,084	65,175,915	65,608,367	65,736,235	66,057,197	66,188,760	66,351,043
20	52,066,233	53,844,254	55,306,358	56,168,419	56,834,780	57,322,399	57,930,955	64,026,114	64,482,213	64,845,907	65,324,118	65,663,196	65,946,061	65,795,199
21	54,499,372	55,151,054	55,817,487	56,104,835	57,282,099	57,680,431	58,491,658	63,854,665	64,377,391	64,858,096	64,933,132	65,479,905	65,461,356	65,605,633
22	53,002,234	53,689,249	54,689,980	55,709,870	56,393,115	56,890,668	57,461,711	63,953,011	64,210,926	64,748,681	64,882,441	65,007,446	65,115,475	65,219,898
23	54,475,618	54,209,390	55,114,979	55,800,549	56,745,566	56,890,699	57,313,138	63,693,211	64,323,138	64,361,084	64,798,794	64,818,666	65,024,302	64,992,457
24	52,696,844	53,885,907	54,088,030	55,057,040	55,463,131	56,208,303	56,444,958	63,400,561	63,820,771	63,700,022	64,438,090	64,258,752	64,141,101	64,186,891
25	50,318,283	51,269,205	52,481,347	53,465,984	53,891,662	54,781,920	54,713,197	63,232,652	63,129,362	63,259,586	63,725,082	63,910,320	63,867,932	63,988,413
26	50,229,788	51,354,758	52,362,605	53,242,078	53,893,195	54,326,699	54,530,116	61,826,182	61,893,288	62,274,584	62,752,607	62,746,811	62,887,121	63,116,564
27	52,206,436	52,688,837	53,341,050	53,769,840	54,270,899	54,576,546	55,060,048	60,352,198	60,726,965	60,923,515	61,131,020	60,963,222	61,265,266	61,663,726
28	49,682,706	50,249,019	51,876,101	52,495,005	52,542,959	53,202,973	53,174,103	60,534,641	60,288,939	60,079,857	59,975,267	60,096,463	59,981,804	60,079,474
29	49,857,350	50,055,633	51,690,368	51,477,619	51,245,707	51,586,655	52,023,933	61,446,881	60,521,500	59,891,967	59,610,978	59,336,539	59,242,416	59,109,188
30	48,055,193	47,429,451	48,033,530	48,151,423	48,174,882	48,183,182	48,215,800	59,390,999	57,149,284	56,340,637	55,831,719	55,327,710	54,997,345	54,953,972
31	39,251,259	39,137,018	39,388,512	39,273,372	39,230,840	39,250,211	39,423,466	47,469,654	45,842,988	45,103,473	44,628,970	44,396,335	44,093,542	44,099,541
Avg Context	52,167,024	53,402,932	54,503,993	55,212,777	56,103,781	56,535,303	56,929,119	63,427,271	63,783,267	64,121,472	64,304,913	64,524,716	64,580,953	64,670,137
32	50,487,094	49,193,302	48,748,380	48,353,375	48,143,528	47,788,231	47,682,812	73,626,535	69,054,469	66,707,831	64,778,405	63,608,111	62,438,327	62,279,482
33	61,392,551	59,894,701	58,602,818	58,523,094	57,350,370	57,349,346	56,865,365	127,465,382	117,081,737	110,181,057	106,211,754	102,176,536	100,298,529	98,864,531
34	61,328,405	61,880,320	61,897,331	61,603,725	61,037,948	60,910,761	61,148,637	183,840,986	178,535,599	174,318,924	169,430,877	164,785,135	162,605,626	161,446,058
35	62,108,896	64,859,473	66,373,167	66,487,737	66,963,760	67,607,923	67,185,378	207,065,353	212,630,432	212,924,641	212,313,484	210,389,783	209,373,568	209,264,567
36	64,491,225	66,128,068	68,573,633	68,479,538	69,624,085	69,468,930	70,076,197	107,042,066	102,720,927	98,446,681	95,544,431	91,884,559	89,921,371	89,176,573
37	64,176,596	67,738,610	71,189,574	72,648,126	74,649,180	75,348,320	76,011,136	78,418,713	75,370,613	73,491,350	72,095,517	70,419,121	69,798,088	69,480,522
38	83,023,501	86,268,700	92,242,626	95,541,179	100,137,087	103,141,517	104,996,137	75,331,828	73,633,103	72,372,719	71,451,162	70,344,087	69,870,019	69,555,251
39	114,997,010	123,345,929	132,119,453	137,345,936	144,729,841	148,911,774	152,506,274	72,416,590	71,374,676	70,711,375	69,920,343	69,170,308	68,691,759	68,547,626
40	72,923,797	75,060,147	76,865,835	77,568,398	78,566,453	79,311,885	80,082,621	73,411,371	72,849,906	71,665,061	71,318,529	70,698,319	70,306,675	70,168,524
41	62,139,748	64,712,748	65,532,787	65,890,626	66,512,689	66,430,631	67,029,855	73,147,695	72,602,565	71,681,923	71,271,172	70,741,320	70,453,928	70,443,127
42	61,518,278	63,119,965	63,331,920	63,642,209	64,136,809	63,861,430	63,990,245	72,618,045	71,781,987	71,440,631	71,063,281	70,644,161	70,394,990	70,299,681
43	58,893,899	59,747,917	60,630,537	60,111,792	60,897,906	60,413,416	60,685,337	72,647,660	71,716,108	71,179,077	70,986,376	70,431,462	70,058,904	70,263,439
44	57,229,696	57,312,370	57,433,177	58,104,220	57,890,592	58,313,878	58,439,482	72,211,146	71,314,701	70,977,086	70,897,382	70,417,594	70,322,615	70,408,343
45	55,873,454	56,139,949	56,811,592	56,589,596	57,622,149	57,870,333	58,004,104	72,853,238	71,729,416	71,312,109	71,062,683	70,958,396	70,941,043	70,812,855
46	54,966,654	55,910,428	56,508,715	56,557,283	57,408,451	57,383,704	58,025,074	73,234,042	72,148,832	71,787,141	71,388,367	71,441,543	70,955,811	71,090,486
47	55,258,100	55,540,073	56,121,577	56,868,270	57,182,726	57,490,540	58,014,824	71,452,519	70,861,810	70,783,165	70,262,176	70,511,868	70,420,548	70,296,965
48	53,806,444	54,635,210	54,818,474	55,121,630	55,692,510	56,391,570	56,482,717	65,034,377	65,452,587	65,749,667	66,046,086	66,391,820	66,416,952	66,424,818
Avg Continuation	64,389,138	65,969,877	67,517,741	68,202,161	69,326,240	69,882,011	70,425,070	92,459,856	90,638,792	89,160,614	88,002,472	86,765,537	86,074,632	85,813,109
Avg Sentence	56,495,689	57,853,725	59,113,028	59,813,184	60,786,735	61,262,262	61,708,935	73,709,645	73,294,599	72,989,502	72,697,798	72,401,674	72,193,298	72,158,273

Table 6: Bi-gram Statistics for half-memorized and quarter-memorized content.