UNCOVERING LATENT MEMORIES IN LARGE LANGUAGE MODELS

Sunny Duan
Brain and Cognitive Sciences
MIT
sunnyd@mit.edu

Mikail Khona
Physics
MIT
mikail@mit.edu

Abhiram Iyer
EECS
MIT
abiyer@mit.edu

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

Ila Rani Fiete
Brain and Cognitive Sciences
MIT
fiete@mit.edu

ABSTRACT

Frontier AI systems are making transformative impacts across society, but such benefits are not without costs: models trained on web-scale datasets containing personal and private data raise profound concerns about data privacy and misuse. Language models are trained on extensive corpora including potentially sensitive or proprietary information, and the risk of data leakage, where the model response reveals pieces of such information, remains inadequately understood. Prior work has demonstrated that sequence complexity and the number of repetitions are the primary drivers of memorization. In this work, we examine the most vulnerable class of data: highly complex sequences that are presented only once during training. These sequences, often containing the most sensitive information, pose a considerable risk if memorized. By analyzing the progression of memorization for these sequences throughout training, we uncover a striking observation: many memorized sequences persist in the model's memory, exhibiting resistance to catastrophic forgetting even after just one encounter. Surprisingly, these sequences may not appear memorized immediately after their first exposure but can later be "uncovered" during training, *even in the absence of subsequent exposures* – a phenomenon we call "latent memorization." Latent memorization presents a serious challenge for data privacy, as sequences that seem hidden at the final checkpoint of a model may still be easily recoverable. We demonstrate how these hidden sequences can be revealed through random weight perturbations, and we introduce a diagnostic test based on cross-entropy loss to accurately identify latent memorized sequences.

1 INTRODUCTION

Frontier AI models are trained on vast web-scale datasets (Touvron et al., 2023; Gemini Team et al., 2023; OpenAI et al., 2023; Brown et al., 2020). The sizes of these pretraining corpora enable fluency, knowledge about various domains (AlKhamissi et al., 2022; Guu et al., 2020), and the ability to perform in-context learning (Brown et al., 2020). However, these datasets often include proprietary, copyrighted, or otherwise private information (Smith et al., 2023; Karamolegkou et al., 2023; Bordt et al., 2024; Duan et al., 2024; Staab et al., 2023; Shi et al., 2023; Tang et al., 2023; Zanella-Béguelin et al., 2019), which is problematic because LLMs have been shown to possess a vast capacity for detailed memorization. Specifically, with appropriate prompting, LLMs can regurgitate verbatim text from their training corpora.

Prior work has found that even sequences encountered early in training can be extracted from the model, long after they have been encountered (Biderman et al., 2023b). One possible cause of this is that memorized sequences appear multiple times within the corpus, allowing the network to reinforce and store this data in its weights. Our findings confirm that repeated sequences constitute the majority

of the memorized content. However, we also find many sequences which are encountered only once during training but are memorized by the model and persist in the model's memory throughout the training process. Many of these memories may seem forgotten during certain stages of training but are later recalled without additional exposure, indicating they remain encoded in the model's weights. These 'latent' memories present significant challenges, as they are not easily detected by current memorization metrics, raising the question of how to effectively identify and quantify memorized training data in large language models.

1.1 CONTRIBUTIONS

This work provides significant insights into the dynamics and mechanics of memorization in language models during pretraining, contributing to the broader understanding of data privacy and security within machine learning. Our primary contributions are as follows:

- **Quantification of Memorization Susceptibility**: We systematically evaluate how the statistical characteristics of training data, specifically sequence complexity and repetition, influence the likelihood of memorization in language models. Our findings demonstrate that the probability of memorizing a sequence scales logarithmically with its repetition in the training data as well as the complexity of the sequence under consideration. These results extends prior work characterizing which sequences become memorized (Prashanth et al., 2024; Tirumala et al., 2022).
- **Stationarity of Memorized Sequences**: By analyzing how memorization changes throughout training, we discover that the memorization status of sequences remains largely stationary after initial exposure, despite not being re-encountered. We find that for many sequences, memorized sequences may disappear and re-appear in the model's output without repeated exposure. This indicates a fundamentally persistent property of the memory, revealing how the state of memorized sequences is preserved and how subsequent training only modifies the model output.
- **Latent Memorization and Recovery**: We identify the presence of "latent" memorized sequences, which are not evident at certain checkpoints but can be uncovered later in training or through controlled perturbations. Our experimental results show that adding random Gaussian noise to model parameters can recover these latent memorized sequences, supporting the hypothesis that further training acts as random additive noise rather than fundamentally altering the memorization state.
- **Development of a Diagnostic Test**: We propose a novel diagnostic test for uncovering latent memorized sequences by analyzing their cross-entropy loss. This test provides a practical tool for detecting and mitigating potential data leakage in deployed language models.

Our study underscores the risks associated with data leakage in language models, emphasizing the need for more robust mechanisms to ensure data privacy. The persistence of memorized sequences poses a challenge for the prevention of data leakage. By characterizing the nature of memorization as well as the nature of these latent memorized sequences, we elucidate possible mechanisms of how sequences become memorized and offer practical solutions for mitigating data privacy risks, and developing safer and more trustworthy models.

2 METHODOLOGY

2.1 PROPERTIES OF PRETRAINING DATA RELEVANT TO MEMORIZATION: REPEATS AND COMPLEXITY

Previous studies have identified that the number of repeats of a sequence affects whether it will be memorized, with more frequently occurring strings being more likely to be memorized (Carlini et al., 2020; Razeghi et al., 2022; Biderman et al., 2023a). Consequently, a starting property to measure is the **number of repeats** of specific strings in the pretraining corpus.

In our work, we also consider a second and newer property: the **complexity** of specific strings. Our decision to do so is motivated by previous studies (Carlini et al., 2020) which identified a prevalent class of easily memorized data: simple sequences composed of repeated patterns, numbers, or other

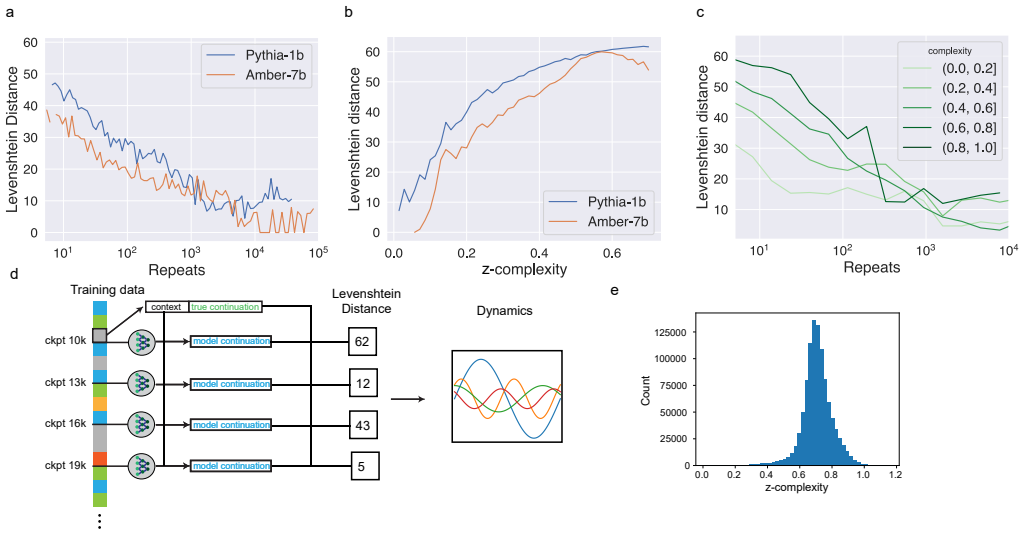

Figure 1: **Data statistics and the probability of memorization a.** Plot of average kl-LD as a function of the number of times the sequence is repeated in the dataset for Pythia-1b and Amber-7b **b.** Average kl-LD as a function of the Z-complexity of the sequence. **c.** Relationship between kl-LD and repeats for different complexity levels. **c.** Schematic of pipeline for analyzing memorization over time. All samples were selected from early on in training. Various model checkpoints are selected and evaluated to determine if early training sequences are still memorized. The changes in k-LD over time are used in our analysis to evaluate how the memorization of these sequences changes throughout training. **d** Distribution of z-complexity over all of the data.

straightforward patterns. While models readily learn these samples, they often lack substantive content and are unlikely to represent sensitive information. Thus, it is important to distinguish between memorization of these trivial sequences from more complex and informative sequences.

In order to quantitatively measure the complexity of specific strings, we use modern compression algorithms to determine the extent to which sequences have a smaller description than the original sequence. To calculate the complexity of a sequence we define a metric, z-complexity, which is the ratio between the compressed sequence length to the original sequence length. This metric contains values from 0 to 1 and is efficiently computable using the `zlib` package in Python.

2.2 Quantifying Memorization

Several different definitions have been put forward to quantify memorization in language models. Intuitively, a memorized sequence is a training sequence which can be reproduced given the right context (Carlini et al., 2022; Schwarzschild et al., 2024). One popular definition of memorization is kl**-memorization** (Carlini et al., 2022). kl-memorization is evaluated by considering a sequence of length $k + l$. The first k tokens are presented to the model as context. The model is used to generate a continuation of length l via greedy (i.e., temperature $= 0$ decoding). The model's continuation is compared to the "true" continuation, and a sequence is said to be kl memorized if the model's output exactly matches the true continuation.

While undoubtedly useful, we introduce and utilize a new metric for measuring memorization. kl-memorization is an overly strict such that even a single-token deviation from the true continuation may cause us to misclassify a sequence as forgotten; in many cases, the model may make small errors such as inserting or modifying a single token. We identified and provide several examples in Table 1. In order to to be more robust to small changes in the learned sequence, we propose a modification of kl-memorization by introducing **kl-Levenshtein distance (kl-LD)**.

Definition 2.1 (kl-LD distance)**.** Let $S = (s_1, s_2, \ldots, s_n)$ be a sequence of tokens. We denote the first k tokens as the context $C = (s_1, \ldots, s_k)$ and the last l tokens as the target $T = (s_{k+1}, \ldots, s_{k+l})$. The

model is provided these context tokens and produces a predicted continuation of $\hat{T} = (m_1, \ldots, m_l)$. We define the kl-LD distance as the Levenshtein distance (Levenshtein, 1965) between sequences T and $\hat{T}$ where the Levenshtein distance is the minimum number of (token) insertions, deletions or substitutions that must be performed on T to obtain $\hat{T}$.

We find that this is a natural measure of memorization which also provides a range of values to provide more granular insight into the strength of the model's memory, akin to how continuous metrics have been used to improve evaluations of language models (Schaeffer et al., 2023; 2024). Throughout this study, we set $k = 32$ and compare the continuation of the model with the original sequence by computing the Levenshtein distance between the next 64 tokens.

Table 1: Model continuations at various stages in training for a few selected sequences which were complex and encountered only once during training. Minimum edits are highlighted such that character edits are highlighted in orange, deletions are highlighted in red and new characters are highlighted in green.

Context	True Continuation	Checkpoint 10000	Checkpoint 15000	Checkpoint 19000
992> por favor ayudenme para instalar DBDesigner <him> hay <BluesKaj>!es | Guest17992	<ubottu> Guest17992: En la mayorÃxada de canales Ubuntu se comunica en inglÃ©s. Para ayuda en EspaÃ±ol, por favor entre en los canales #ubuntu-es o #kubuntu-es. <BluesKaj>!es |	<ubottu> Guest17992: En la mayorÃ a de canales de Ubuntu se comunica sÃ³lo en inglÃ©s. Para busca ayuda en EspaÃ±ol, por favor entrera en los canales #ubuntu-es o #kubuntu-es. <BluesKaj>! es |	<ubottu> Guest17992: En la mayorÃ a de canales de Ubuntu se habla sÃ³lo en inglÃ©s. Si busca ayuda en espaÃ±ol o charlar entra en el canal #ubuntu-es. Escribe "/join #ubuntu-es"	<ubottu> Guest17992: En la mayorÃ a de los canales de Ubuntu, se habla sÃ³lo en inglÃ©s. Si busca ayuda en espaÃ±ol entre al lar entra en el canal #ubuntu-es, escriba "/join #ubuntu-es" |
, findings, beliefs, or experiences on those topics or products. The views and opinions expressed on CateTheOkay.com are purely my own. Any product	claim, statistic, quote or other representation about a product or service should be verified with the manufacturer, provider or party in question. CateTheOkay.com doesn't contain any content which might present a conflict of interest.	claim, statistic, quote or other representation about a product or service should be verified with the manufacturer or provider. Comments. I have a question. I have a friend who is a teacher and she is a teacher. She is a teacher and she is a student. She is a student and she is a	claim, statistic, quote or other representation about a product or service should be verified with the manufacturer or provider or party in question. CateTheOkay.com is not affiliated with, endorsed by, or sponsored by the Coca-Cola Company. CateTheOkay.com is not affiliated with, endorsed by,	claim, statistic, quote or other representation about a product or service should be verified with the manufacturer or provider or party in question. I am not a doctor, pharmacist, or registered dietitian. I am not a registered dietitian. I am not a registered dietitian. I am not a registered dietitian. I am

2.2.1 Analyzing Repeated Samples

In this study, we seek to understand both how repeated encounters of a sequence during training drives memorization and also how sequences which are encountered only once are retained by the model. To this end, we analyze where training sequences are repeated throughout the course of training. In our study, we focus on the l portion of the sequence. For this study, we fixed l to 64 tokens. Given a target sequence, we compare the target sequence with all of the training sequences which were presented to the model during the period of training under consideration. We compute the largest subsequence match between the target and every individual training example and call a training example a "repeat" if there was a sub-sequence match of length 30 or longer.

2.3 Language Models

In this study, we largely focused on the Pythia 1B language model (Biderman et al., 2023a), which was trained on 300B tokens from the Pile (Gao et al., 2020). For selected experiments, to ensure our results hold on other language models, we reproduced our results using a second model, Amber-7B (Liu et al., 2023). We selected these two models as they were large, high performing models complete with fully reproducible data sequences and frequent checkpoints. As in previous works (Biderman et al., 2023a), all experiments were run with the models run with half precision (`float16`) and temperature 0.

2.4 Datasets

In this work, we use the deduplicated versions of the Pile (Gao et al., 2020) as well as the Amber dataset which is a combination of the RedPajama V1, RefinedWeb and StarCoderData datasets (Computer, 2023; Li et al., 2023; Penedo et al., 2023), all of which employ deduplication. In the early part of our work we also employ the standard Pile which did not use de-duplication in order to

observe the effects of repeated exposure on memorization. In the latter part of our work we focus on the deduplicated versions of the datasets in order to eliminate the influence of repeated exposure on our analysis of memorization.

2.4.1 LANGUAGE MODEL CHECKPOINTS

In our analysis, we used checkpoints from every 3k training steps between from step 10k-43k in Pythia-1B and every 10 checkpoints of Amber-7B, corresponding to roughly 1.7 million training examples between revision 100 to 350. These selections were checkpoints from each model which represented a sizable portion of training. These were chosen to be offset from the beginning of training to avoid artifacts or initial transients from random initialization, learning rate warmup and other peculiarities from initial phases of training.

3 EXPERIMENTAL RESULTS

3.1 DATA STATISTICS PREDICT MEMORIZATION

We analyze two primary drivers of memorization during training: sequence complexity, and the number of repetitions. Previous work showed that the probability a training string can be extracted from a model is related to the model size and number of repetitions (Carlini et al., 2020); we find that this relationship is true in the models we analyzed as well (Figure 1a). Additionally, we found that the complexity of the string itself was a strong predictor of whether it would be memorized (Figure 1b): strings with smaller z-complexity had smaller kl-Levenshtein distance (kl-LD), meaning simpler strings are more easily memorized. Interestingly, recent work showed that pretraining language models on data with lower z-complexity causes the training losses to decrease more rapidly (Pandey, 2024); our results here suggest an explanatory mechanism: with more compressible data, the model can memorize the data more quickly. Furthermore, we found that for strings of different complexity exhibited different memorization curves (Figure 1c), whereby lower complexity strings were more easily memorized with fewer repeats. Both of these factors influenced the memorization probability with a log-linear relationship. Since highly sensitive information is likely contained in complex and rare training sequences, we focus our efforts on these sequences. **In the rest of this work, we restrict our analys to sequences which are presented once and have high complexity** (> 0.8).

3.2 DYNAMICS OF MEMORIZATION

In order to produce a more complete picture of how successive training affects the state of memorized sequences within our model, we analyze how the kl-LD changes throughout the course of training for individual sequences. We select sequences early on in training and evaluate how the memorization status of these sequences evolves throughout training (Figure 1a). In this section we utilize the deduplicated version of the Pile dataset (Gao et al., 2020) as well as the Amber LLM360 dataset which also uses deduplication in order to remove the effects of repeated exposures. We filter these sequences so that only sequences which have a z-complexity of 0.8 or higher are included in our analysis. Additionally, we employed our own duplication detection scheme which eliminated sequences which had a sub-sequence match of length 30 or longer and rerun our analysis using these sequences (Figure S13).

Surprisingly, we find that the memorization status of a sequence is largely stationary throughout training. After the initial checkpoint, the kl-LD of the sequences fluctuate (Figure 3) but do so in a way which is stationary across training (Figure 2a). This is consistent across both Pythia-1b and Amber-7b models. This is reflected in the individual trajectories, and also in the overall mean of the population which shows no clear trend as training progresses. Furthermore, unlike a random walk, we see that the variance of the does not grow over time, but remains fixed. We can quantify this by running a variance ratio test A.2, where the variance of a random walk is expected to grow linearly. We can reject the null hypothesis that our data is generated from a random walk with $p < 10^{-8}$ for the samples drawn from Pythia-1b and Amber-7b. This is indicative of a mean reversion tendency of the dynamics and demonstrate the stability of the memories within the model weights. Additionally, we observe that the changes in the kl-LD between consecutive checkpoints are symmetric (Figure 2c) and roughly follow a laplace distribution (Figure 3). This again confirms the counter-intuitive property of sequences to become memorized as often as they are forgotten. Notably, the model is able

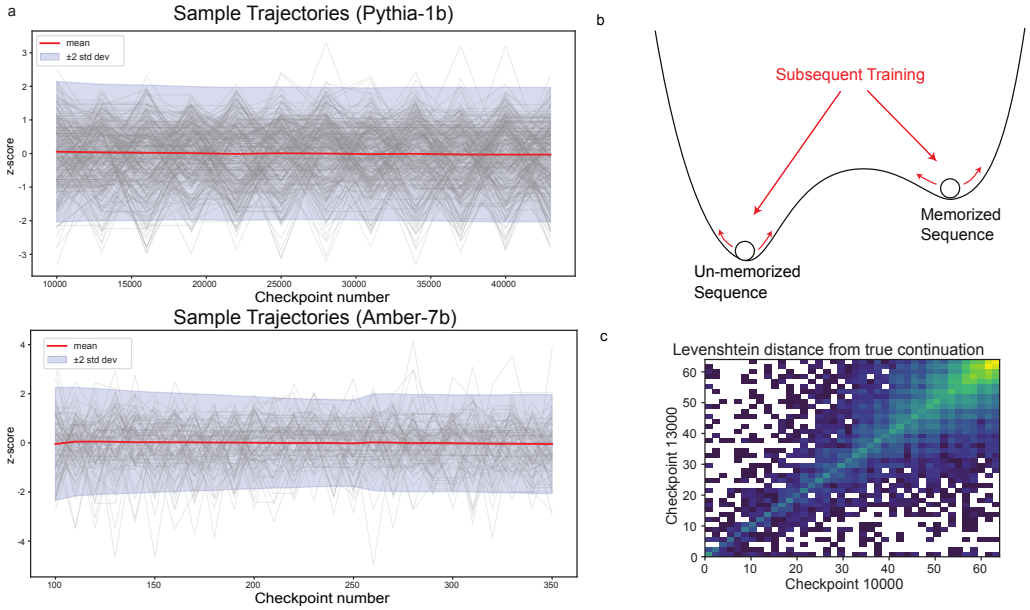

Figure 2: a. Visualization of individual samples and the change in the memorized length during training. Grey lines are subsampled single sequence trajectories throughout training. Each sequence was normalized such that the distribution of memorization lengths was mean 0 and variance 1. The red line denotes the mean and shaded area denotes region of two standard deviations of the kl-LD of all sequences at a single point in time. Notably, the distribution at each timestep is the same for all checkpoints. This is in contrast to both the expected exponential decay behavior exhibited by models which experience catastrophic forgetting as well as the linear growth of variance which is expected of processes exhibiting random walk behavior. b. Conceptual schematic of how memorized sequences may be stabilized during training in order to resist the interference from weight changes caused by subsequent training. c. Joint distribution of kl-LD during checkpoints 10k and 13k. Color is the log of the number of sequences in each bin. The vast majority of sequences are not memorized in either checkpoint.

to recall memories which, at one point in time, appeared to be forgotten, despite never encountering that sequence again.

The stationarity of the memorization status of these sequences indicates that the memorized sequence is stable throughout time, but this is in conflict with the fact that the model weights are constantly evolving. This stability in the presence of noise is indicative of a stabilizing mechanism by which the encoding of the sequence memory is preserved by some restorative process illustrated in Figure 2b where the memorized sequence becomes a stable fixed point in the weight space of the model under training dynamics. Since this is not true of all sequences, but only the few which exhibit this persistent memorization, it may point to a phase transition that occurs when the sequence is first encountered.

3.3 Latent Memorization and Recovery of Latent Memories

Since some sequences exhibited seemingly random variations in their memorization state across different checkpoints, we hypothesize that these sequences remain memorized but are not be visible at a given checkpoint and are "latent" memorized. Indeed, we found many sequences which were not memorized at the initial checkpoint (10k) but exhibited memorization by checkpoint 19k (Table 1).

For these sequences, the nature of the random changes shown in Figure 2a indicate the form of a random walk. We hypothesize that the process of training in frontier AI models acts as random noise on the weights with respect to the memory of the sequence. Thus, simply perturbing the weights with random noise should produce similar effects as training.

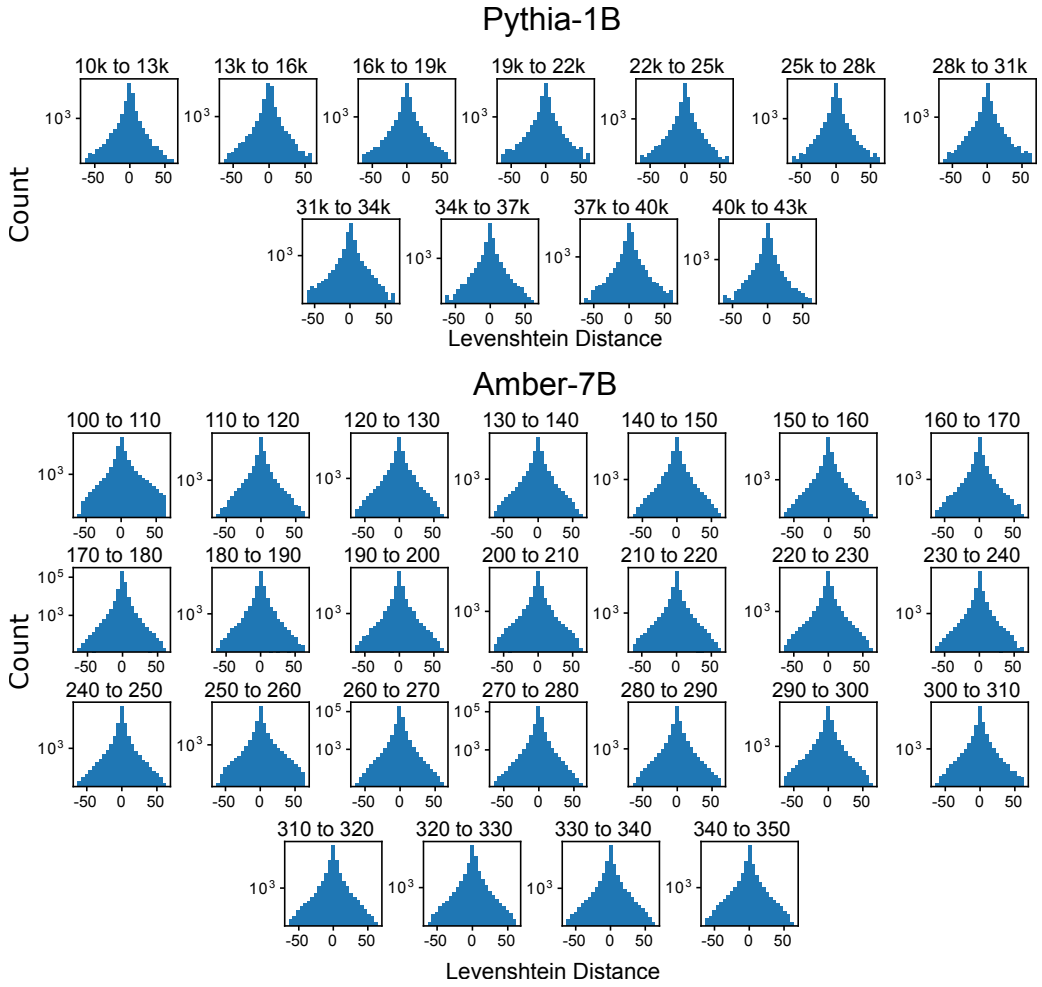

Figure 3: Visualization of how memorization status, measured by kl-LD, changes between consecutive checkpoints for both Pythia-1B and Amber-7B.

We find that this prediction is true. We randomly perturb the model weights by adding a small amount of random gaussian noise (of magnitude 2×10^{-3}) to each of the weight parameters. We repeat this process 200 times and find the perturbation which yields the lowest kl-LD (Figure 4a). Notably, in the high dimensional weight space, it is difficult to reproduce arbitrary sequences using random weight perturbations, thus the recovery of memorized sequences must be due to intrinsic factors of how the memory is encoded in the weights.

We find that sequences which were "latent" memorized are able to be recovered using random perturbation (Figure 4bc). In contrast, sequences which were not memorized during the period of consideration could not be recovered. As a control, we also selected sequences which were not presented to the model yet, and observed that their distributions closely matched those which were encountered by not memorized by the model (Figure 4b). Furthermore, we found that the perturbations yielded memorization patterns which closely matched that of the model at a later point in training. These observations support the view that with respect to a memorized sequence, subsequent training acts similar to random noise perturbations to the model weights.

As an additional control, we attempted to recover the latent memorized sequences by sampling from the model at different temperatures. Since these sequences are stored in the model weights, it may be the case that the model would simply reproduce the target sequences if prompted enough times. We tried four different temperatures and sampled 200 different sequence continuations from each

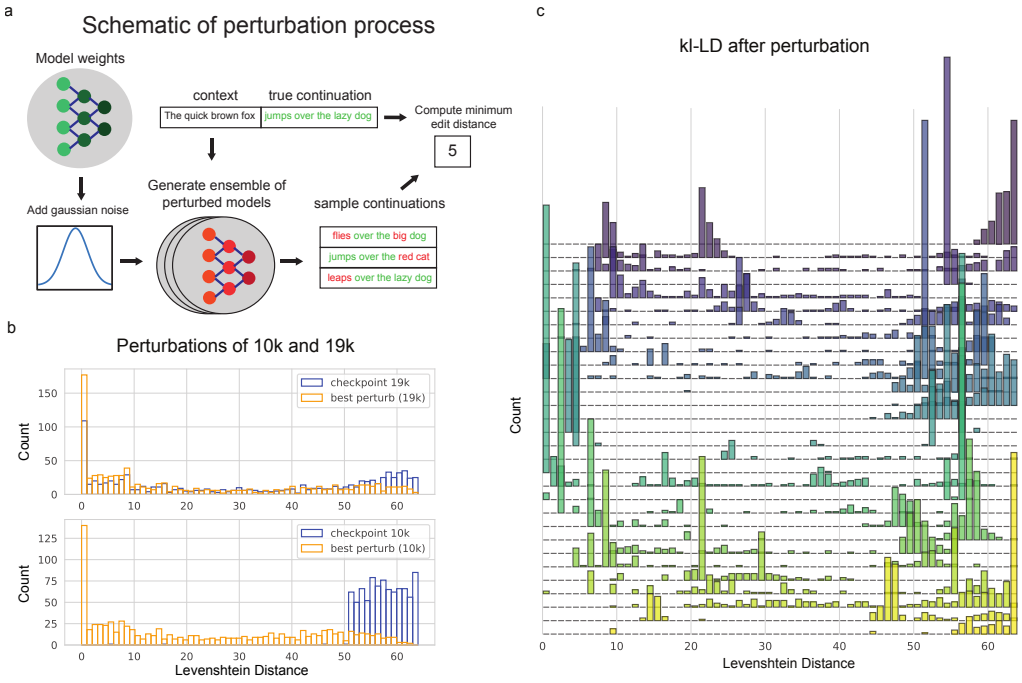

Figure 4: **a**. Schematic of how model weights are perturbed in order to extract memories from the model. The same context is given to 200 perturbed modesl and the best continuation is chosen. **b**. Comparison of the distribution of best achievable kl-LD by perturbing the model weights. Data points were selected such that they were un-memorized (kl-LD > 50) at 10k but we're memorized (kl-LD < 10) at some point during the next 10k training steps. Top panel is the histogram of the perturbations of the checkpoint at 19k and bottom is 10k. Notably, the perturbations cause the 10k model distances to match the distribution of the 19k model, and perturbing the 19k model does not have a significant effect. This is indicative of how model training mimics random noise with respect to the memorization status of the sequences. **c.** Visualization of the Levenshtein distances from the target for various weight perturbations. Each row is a single sequence, and the heights of the bars correspond to the number of perturbations which resulted in a Levenshtein distance of that corresponding bin.

of the temperatures. We find that the this method fails to recover the latent memories that weight perturbation was able to produce (Figure 5b).

"Latent" memorized sequences pose a significant risk for leakage since they are not easily detectable from evaluating kl-memorization of those sequences. To this end, we discovered that these "latent" memorized sequences had significantly lower cross entropy loss when evaluated by the model (Figure 5c), thus simply evaluating the likelihood of those sequences using the trained model is a natural diagnostic for detecting these "latent" memorized sequences.

3.4 Related Work

Extracting memorized sequences from language models is an area of high interest. Early work established that it was possible to extract sensitive data including phone numbers, URLs and personal information from trained language models (Carlini et al., 2020). Other studies injected canaries to determine what aspects of the training process contributed to whether a sequence is extractable (Henderson et al., 2017)(Thakkar et al., 2020). More recent work have extended this to investigate how these properties scale with model size and data statistics (Carlini et al., 2022). This has motivated the use of deduplication, which in addition to reducing the chance of data leakage (Kandpal et al., 2022), also has been shown to improve sample efficiency and improve evaluation (Lee et al., 2021).

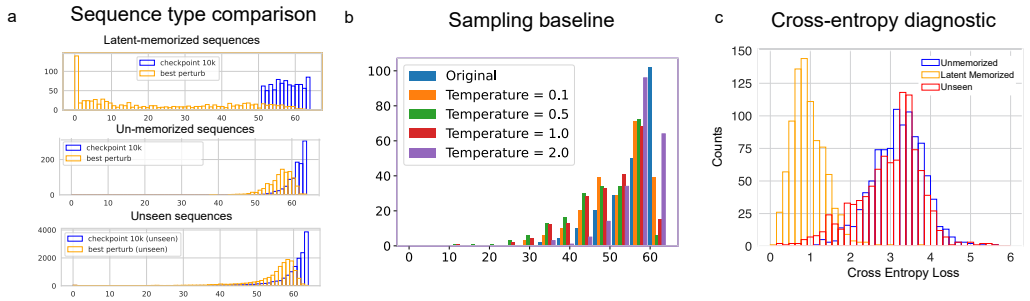

Figure 5: **a.** Comparison of using perturbations to evoke a target sequence for three different classes of sequences. In the top panel, we examine the sequences which are "latent" memorized. In the middle panel, we find sequences which weren't memorized during training and in the bottom panel, we analyze sequences which were encountered later in training but were not encountered by the model. We not that perturbing the weights is only able to evoke sequences which are "latent" memorized. **b.** Attempts at evoking latent memories using 200 samples at various temperatures. None of the temperatures were able to reliably recover the latent memories. **c.** Comparison of the cross entropy losses of sequences separated into the three different classes of sequences analyzed in b. The cross entropy losses of "latent" memorized sequences are much lower.

The definition of memorization is also still debated and various approaches to quantifying memorization have been made (Zhang et al., 2021; Feldman & Zhang, 2020). A variety of attacks have been designed to extract memorized sequences using designed prompts (Thakkar et al., 2020) and model activation perturbations (Kassem et al., 2024).

More generally, the notion of membership inference has been studied as a way to determine whether a given training example was part of the corpus (Shokri et al., 2016; Mireshghallah et al., 2022; Hisamoto et al., 2019), and these approaches have been applied to language models as well (Duan et al., 2024).

Forgetting has also been studied extensively in neural networks, typically in the context of preventing forgetting. (Kirkpatrick et al., 2017; Zenke et al., 2017; Chen et al., 2020). Studies have also shown that forgetting decreases with model size (Tirumala et al., 2022; Mirzadeh et al., 2021). This work has also been examined in the context of understanding what aspects about a model and the data contribute to forgetting (Toneva et al., 2018)

Finally, there has also been work studying how the training process affects the status of memorization (Tirumala et al., 2022; Prashanth et al., 2024). This work focuses on how parameters of training and size of the model affect the dynamics of training. They find that scaling the model generally leads to less forgetting. In our work, we focus on sequences which counter-intuitively do not obey the forgetting laws presented in this work and expanding on the implications of these persistent "episodic" memories.

4 Conclusion and Limitations

We study how memorization changes throughout training and focused on sequences which occurred only once throughout training. Under these conditions, we find that rather than forgetting these sequences, the model retains them for the duration of training. This stationarity indicates a stability of the memorized sequence in weight space since the training process necessarily modifies the weights which encode the memorized sequences. We test this mechanistic view of how the training process interacts with the memorized sequence by using random weight perturbations to the model weights. These perturbations confirm that sequences which appeared to be forgotten at one point during training, may still be memorized by the model and are able to be uncovered with a small amount of random noise. We concluded by demonstrating a simple diagnostic to distinguish between "latent" memorized sequences and un-memorized sequences.

This study highlights one surprising behavior of frontier AI models and begins to uncover what mechanisms are present in the memorization behavior of these models. Our work suggest a possible mechanism of how memorized strings are sustained throughout training and further experiments are needed to confirm the underlying mechanism. Notably, further testing is required across other frontier AI models which were not considered here. Finally, we propose a mechanistic explanation for this phenomenon which requires further study to explain the cause of these persistent memories.

5 Reproducibility Statement

All code used for this project is available at `https://github.com/sunnyddelight/latent_memorization`.

REFERENCES

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. A review on language models as knowledge bases. *arXiv [cs.CL]*, April 2022.

Stella Biderman, Usvsn Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *arXiv [cs.CL]*, April 2023a.

Stella Biderman, Usvsn Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin G Anthony, Shivanshu Purohit, and Edward Raf. Emergent and predictable memorization in large language models. *Adv. Neural Inf. Process. Syst.*, abs/2304.11158, April 2023b.

Sebastian Bordt, Harsha Nori, Vanessa Rodrigues, Besmira Nushi, and Rich Caruana. Elephants never forget: Memorization and learning of tabular data in large language models. *arXiv [cs.LG]*, April 2024.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv [cs.CL]*, May 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *arXiv [cs.CR]*, December 2020.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv [cs.LG]*, February 2022.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. *arXiv [cs.CL]*, April 2020.

Together Computer. RedPajama: an open dataset for training large language models, October 2023.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv [cs.CL]*, February 2024.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *arXiv [cs.LG]*, August 2020.

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. The pile: An 800GB dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, December 2020.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, Yaguang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, Hyunjeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas

Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M R Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-Yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran

Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Kumar Reddy M Pavan, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, Z J Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, L V Pallavi, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, R Raghavender, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan,

Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, L I N Tian, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser Tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A Choquette-Choo, Yunjie Li, T J Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver

Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, Xianghai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M K Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. Gemini: A family of highly capable multimodal models. *arXiv [cs.CL]*, December 2023.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. *arXiv [cs.CL]*, February 2020.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. Ethical challenges in data-driven dialogue systems. *arXiv [cs.CL]*, November 2017.

Sorami Hisamoto, Matt Post, and Kevin Duh. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *arXiv [cs.LG]*, April 2019.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. *arXiv [cs.CR]*, February 2022.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Sogaard. Copyright violations and large language models. *arXiv*, October 2023.

Aly M Kassem, Omar Mahmoud, Niloofar Mireshghallah, Hyunwoo Kim, Yulia Tsvetkov, Yejin Choi, Sherif Saad, and Santu Rana. Alpaca against vicuna: Using LLMs to uncover memorization of LLMs. *arXiv [cs.CL]*, March 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. U. S. A.*, 114(13):3521–3526, March 2017.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv [cs.CL]*, July 2021.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. StarCoder: may the source be with you! *arXiv [cs.CL]*, May 2023.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P Xing. LLM360: Towards fully transparent open-source LLMs. *arXiv [cs.CL]*, December 2023.

Fatemehsadat Mireshghallah, Kartik Goyal, Archit Uniyal, Taylor Berg-Kirkpatrick, and Reza Shokri. Quantifying privacy risks of masked language models using membership inference attacks. *arXiv [cs.LG]*, March 2022.

Seyed Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Gorur, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. *arXiv [cs.LG]*, October 2021.

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, C J 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. GPT-4 technical report. *arXiv [cs.CL]*, March 2023.

Rohan Pandey. gzip predicts data-dependent scaling laws, 2024.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only. *arXiv [cs.CL]*, June 2023.

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

Yasaman Razeghi, Robert L Logan I V au2, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot reasoning, 2022.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In A Oh, T Naumann, A Globerson, K Saenko, M Hardt, and S Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 55565–55581. Curran Associates, Inc., 2023.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier AI models with scale remained elusive?, 2024.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. Rethinking llm memorization through the lens of adversarial compression. *arXiv preprint arXiv:2404.15146*, 2024.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. *arXiv [cs.CL]*, October 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *arXiv [cs.CR]*, October 2016.

Victoria Smith, Ali Shahin Shamsabadi, Carolyn Ashurst, and Adrian Weller. Identifying and mitigating privacy risks stemming from language models: A survey. *arXiv [cs.CL]*, September 2023.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. *arXiv [cs.AI]*, October 2023.

Ruixiang Tang, Gord Lueck, Rodolfo Quispe, Huseyin A Inan, Janardhan Kulkarni, and Xia Hu. Assessing privacy risks in language models: A case study on summarization tasks. *arXiv [cs.CL]*, October 2023.

Om Thakkar, Swaroop Ramaswamy, Rajiv Mathews, and Françoise Beaufays. Understanding unintended memorization in federated learning. *arXiv [cs.LG]*, June 2020.

Kushal Tirumala, Aram H Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *arXiv [cs.CL]*, May 2022.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv [cs.LG]*, December 2018.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, February 2023.

Santiago Zanella-Béguelin, Lukas Wutschitz, Shruti Tople, Victor Rühle, Andrew Paverd, Olga Ohrimenko, Boris Köpf, and Marc Brockschmidt. Analyzing information leakage of updates to natural language models. *arXiv [cs.LG]*, December 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. *Proc Mach Learn Res*, 70:3987–3995, 2017.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. *arXiv [cs.CL]*, December 2021.

A Appendix / Supplemental Material

A.1 Compute details

All experiments were run on a cluster with access to 16 concurrent a100 GPUs. All of the language models were run using a single GPU and multiple GPUs were used to parallelize the experiments in order to speed up progress. Searching for repeats within the dataset was performed using the library dask, using 64 CPUs distributed in a cluster, each with 32Gb of RAM.

A.2 Variance Ratio Test

Random walks have a hallmark property of linearly increasing variance over time. We can demonstrate statistically that the sequence of memorization lengths does not follow a random walk by conducting a variance ratio test. Given a sample $\{\{X_ij\}_{1\leq i\leq m}\}_{1\leq j\leq n}$ of m sequences of length n, we can calculate the F-statistic by taking the ratio of the variances

$$\frac{\frac{1}{m}\sum_{i=1}^{m}(X_{in}-\bar{X}_{in})^2}{n\frac{1}{m}\sum_{i=1}^{m}(X_{i1}-\bar{X}_{i1})^2}$$

which, for a random walk has an F-distribution with m and m degrees of freedom.

A.3 Licenses

This project used code from the Pythia project Biderman et al. (2023a) released by EleutherAI under the Apache license version 2.0. We also used the Pile dataset Gao et al. (2020) which is released under the MIT license. The Amber model was produced by LLM360, and the code and dataset are both released under Apache 2.0.

A.4 Additional figures

We include figures which were omitted from the main paper. These provide additional details that were not central to the claims made in the paper.

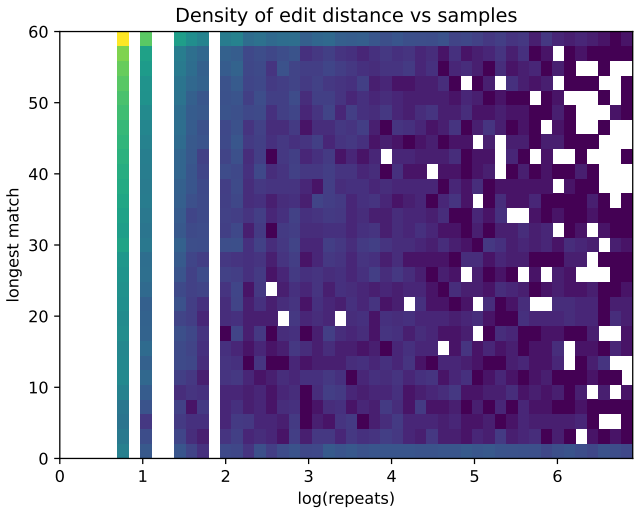

Figure 6: **Histogram of the repeats vs the edit distance** Hue is log density.

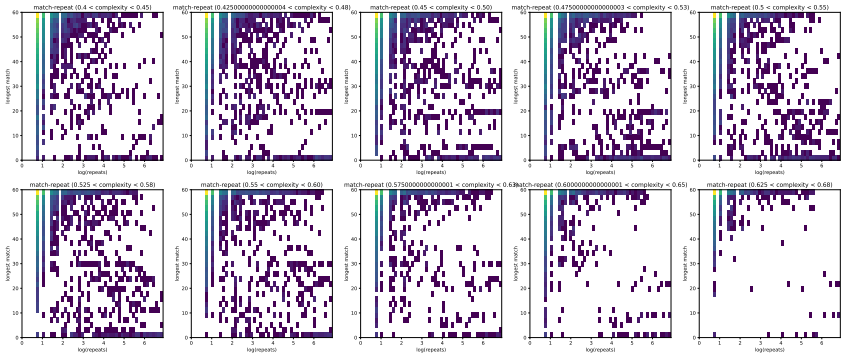

Figure 7: **Histogram of the repeats vs the edit distance split by complexity** Hue is log density.

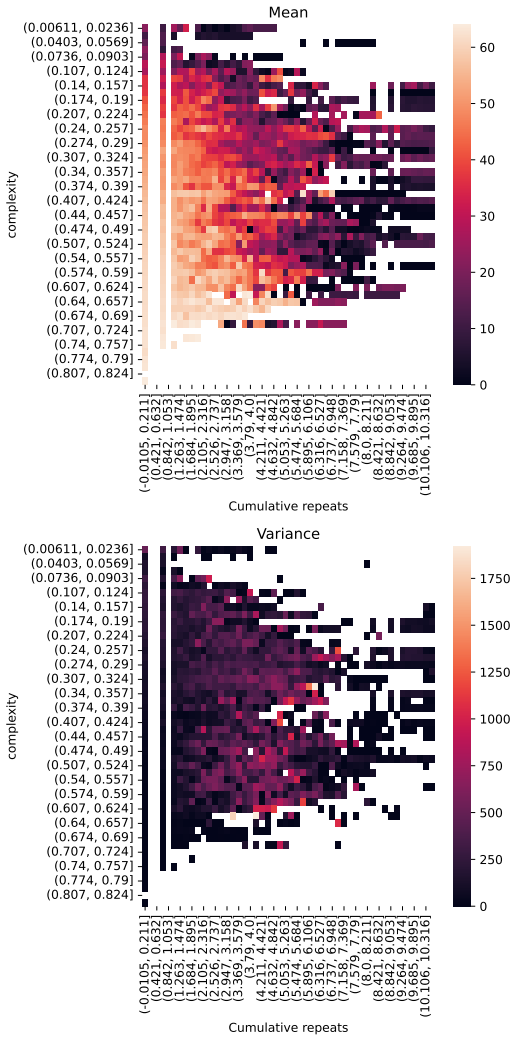

Figure 8: **Average of the kl-LD metric** kl-LD values are binned by number of repeats and complexity and the mean and variance of the samples in those bins are computed and colored.

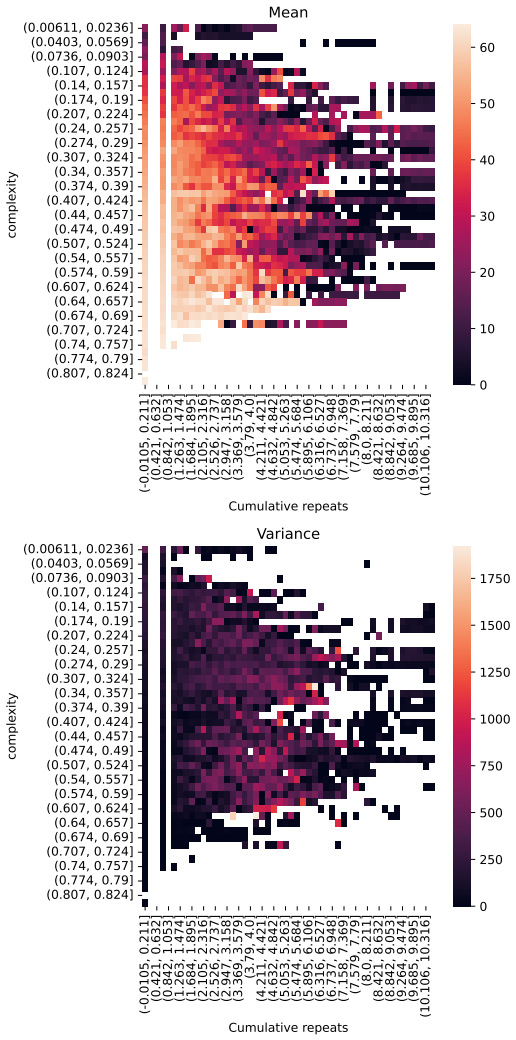

Figure 9: **Average of the kl-LD metric** kl-LD values are binned by number of repeats and complexity and the mean and variance of the samples in those bins are computed and colored.

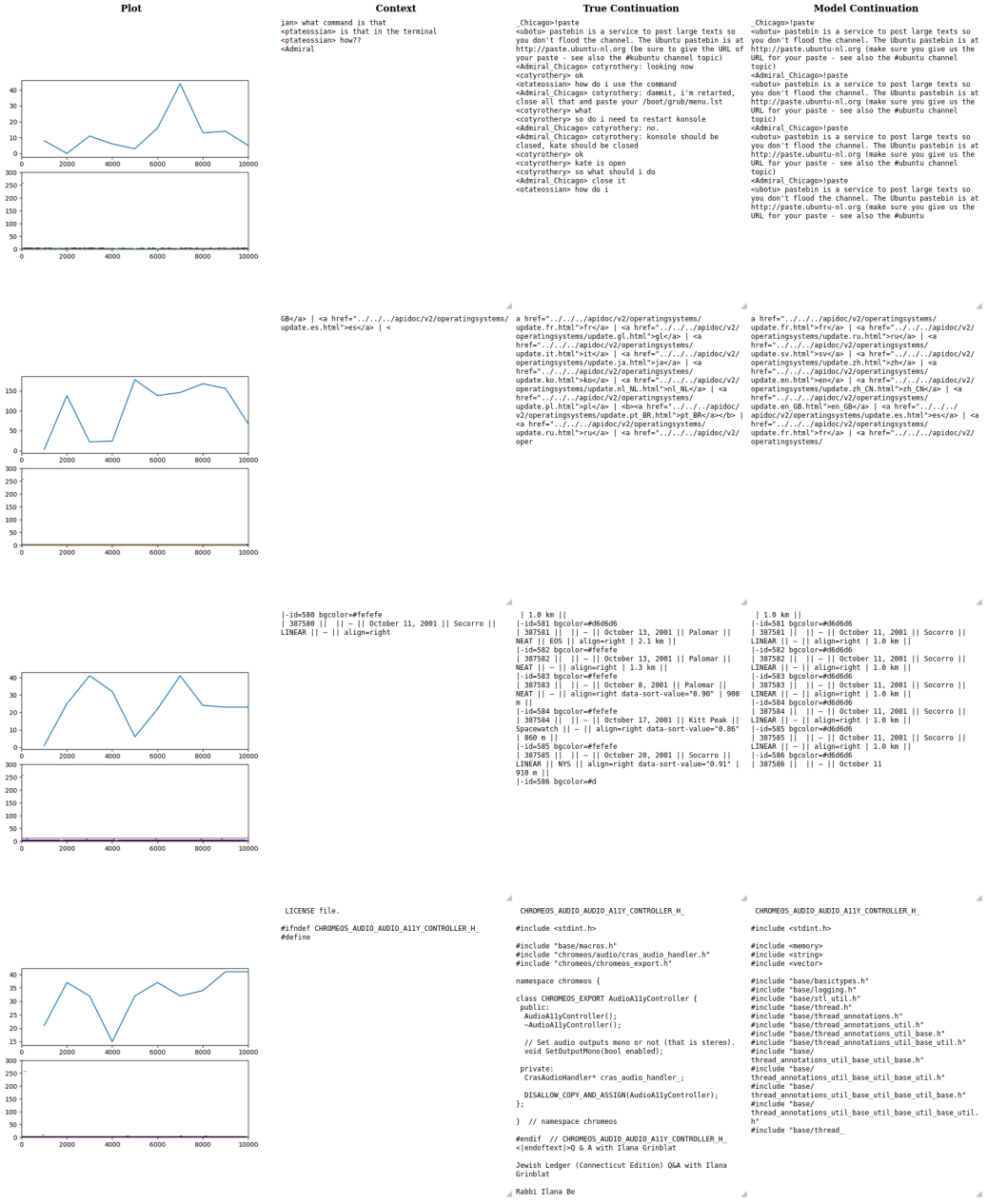

Figure 10: Examples of strings which were seen once during training. Top left plot shows the kl-LD over for different trajectories and bottom left plot is a histogram of when the examples were repeated and at what length with the time on the x axis and the length of the repeat on the y axis. The text of the context, true continuation and model continuation are shown as well.

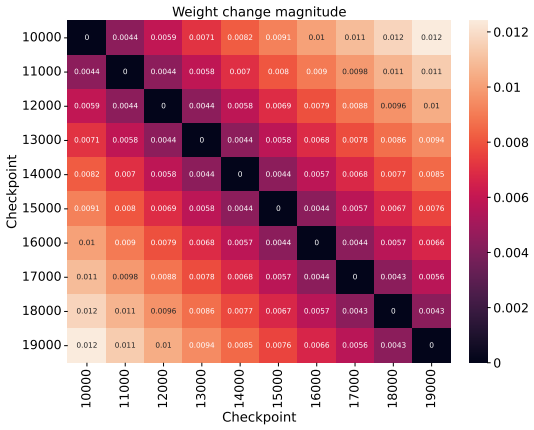

Figure 11: Distribution of weight changes of the model throughout training. Computed as the L2 distance between the flattened model weights at two different times during training

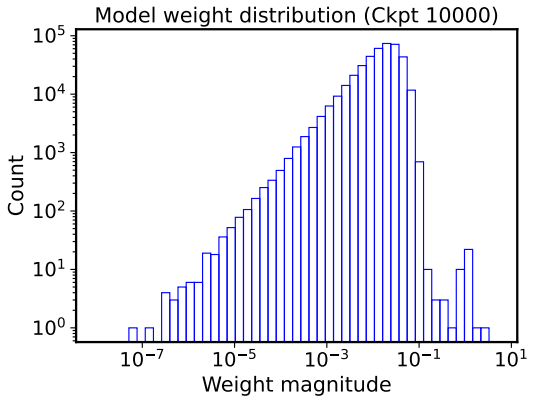

Figure 12: Distribution of weight magnitudes of the trained model at checkpoint 10k

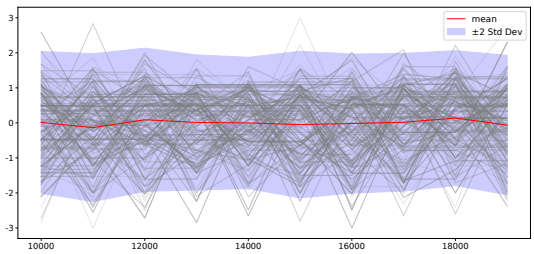

Figure 13: Dynamics of sequences with high complexity (>0.7) filtered by using maximum substring match of 30