Exploring Forgetting in Large Language Model Pre-Training

Anonymous ACL submission

Abstract

In large language models (LLMs), the challenge of catastrophic forgetting remains a formidable obstacle to building an omniscient model. Despite the pioneering research on task-level forgetting in LLM fine-tuning, there is scant focus on sample-level forgetting during this phase, where models often see each datapoint only once. We systematically explore the existence, essence, and measurement of forgetting in LLM pre-training, questioning traditional metrics such as perplexity (PPL) and introducing new metrics to better detect entity memory retention, which is the indicator of forgetting. Taking inspiration from human memory patterns, we propose and refine memory replay techniques to combat the phenomenon of forgetting in LLMs. Extensive evaluations and analyses on forgetting of pre-training could facilitate future research on LLMs.

1 Introduction

Catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990) poses a significant challenge to the development of models that are capable of continuous learning, which is also observed in LLMs. Traditionally, the challenge of **catastrophic forgetting** in neural networks is especially pronounced when models are tasked with retaining knowledge across diverse datasets (Sun et al., 2020; Jin et al., 2021; de Masson D'Autume et al., 2019; Wang et al., 2020; Qin et al., 2022), necessitating a delicate balance between the acquisition of new information and the retention of previously learned knowledge. This issue arises due to the shift in input distribution across different tasks, which leads to the model's inability to remember past information effectively.

Although pioneer efforts have explored the forgetting issue in LLM fine-tuning, which primarily addresses task-specific forgetting, there is a lack of research on finer-grained forgetting *at the sample level* in **pre-training**. Luo et al. (2023), Wang et al. (2023b), and Wu et al. (2024) focused on forgetting in fine-tuning by measuring the performance of new tasks with continual tuning. Other efforts (Tirumala et al., 2022; Biderman et al., 2023a) studied sample-level memorization, where some experiments imply the existence of forgetting in LLM pre-training. Nonetheless, these studies have devoted limited attention to systematically exploring and quantifying the forgetting in pre-training.

Forgetting in LLM pre-training is a critical issue that must be addressed. It is prevalent among current LLMs and significantly affects their performance. Intuitively, after fine-tuning, LLMs may give unsatisfactory replies to queries, even when the necessary information was present in the pre-training data. This indicates forgetting. Despite being easily noticed, measuring this forgetting in pre-training is very challenging. In contrast to works studying forgetting in fine-tuning that measure with specific task-related metrics (e.g., QA accuracy), the pre-training stage is not optimized for specific tasks or datasets. Moreover, the conventional LLM metrics such as perplexity (PPL) are also shown to be insensitive in measuring forgetting in pre-training (Gupta et al., 2023). This raises a pertinent question: (1) *How to correctly recognize and quantify forgetting in pre-training?*

As new metrics emerge to quantify forgetting in pre-training, we draw inspiration from the proven success of episodic memory replay methods in combating forgetting **during dataset shifts**, as shown in (de Masson D'Autume et al., 2019; Wang et al., 2020), and delve into the inquiry: (2) *Can these methods also mitigate forgetting during the pre-training phase?*

Starting from the premise that higher review intensity slows down the forgetting rate in human learning (Loftus, 1985), we notice that traditional episodic memory replay methods for models employ a lower intensity of learning for the replayed samples. This observation prompts the question

of whether models' forgetting behaviors mirror human learning patterns to any extent. With this in mind, we are interested in investigating if increasing the intensity of memory replay could improve the retention in models. We pose the inquiry: (3) *Do models exhibit forgetting patterns that mirror human learning curves? Can leveraging these patterns through intensified memory replay mitigate forgetting during pre-training?*

To address the above questions, we first magnify the forgetting issue by building a didactic scenario, and scrutinize the limitation of conventional metrics (e.g., PPL) in identifying pre-training forgetting. Next, looking deeper into the essence of pre-training forgetting, we conclude that **the recall ability of entity-related information** is one of the most explicit and significant indicator of forgetting during pre-training. Subsequently, we propose three novel entity-related metrics and experimentally confirm the existence of forgetting during pre-training. Within a standard pre-training setting, we present several simple and effective memory replay strategies, demonstrating that our straightforward replay tactics can alleviate the forgetting issue during pre-training.

Finally, we explore the impact of repeatedly learning from replayed samples in a short period. We examine how the metrics of these recently learned samples evolve over the course of further learning. Drawing an analogy to the human memory curve, we explore the impact of short-term, high-frequency learning on the model's memory retention, shedding light on future pre-training designs aimed at mitigating forgetting.

Our contributions are summarized as follows: (1) We systematically explore and quantify the phenomenon of forgetting during pre-training through new entity-focused metrics. (2) We examine the effectiveness of memory replay in reducing pre-training forgetting. (3) We further examine how short-term, high-frequency learning affects the model's memory retention.

2 Related Work

Catastrophic Forgetting in Language Models. Neural networks often experience catastrophic forgetting when learning new tasks, losing old knowledge due to changes in data distribution (McCloskey and Cohen, 1989; Ratcliff, 1990). Various strategies have been proposed to counter this, such as simultaneous training of new and old tasks (Sun et al., 2020), incremental lifelong pre-training with continual learning algorithms (Jin et al., 2021), and the incorporation of episodic memory to handle diverse data distributions (de Masson D'Autume et al., 2019). Other approaches include meta-lifelong frameworks (Wang et al., 2020) and function-preserved model expansion (Qin et al., 2022). However, most of these studies do not deeply explore single data distribution scenarios. Our study uniquely focuses the pre-training phase, offering fresh insights into forgetting.

Example Forgetting and Forgetting During Pre-training. Despite significant research on catastrophic forgetting, there is limited investigation into forgetting within the context of a single task. Toneva et al. (2018) first defined example forgetting, where certain examples are correctly classified and later misclassified during training. Tirumala et al. (2022) explored forgetting dynamics in language models. Biderman et al. (2023a) studied model behavior forecasting, while Gupta et al. (2023) examined warm-up strategies in continual pre-training. However, a detailed formalization of what is forgotten and how it is quantified using metrics has been lacking—this is where our research steps in.

3 Existence of Pre-training Forgetting

3.1 Intuition on Pre-training Forgetting

First, we explore whether, **after pre-trained**, an LLM *exhibits a pattern of decreased performance on earlier samples*, suggesting sample-level forgetting in pre-training. To test this, we take a direct approach: after training, we obtain a checkpoint and then **use this exact checkpoint** to test on samples in the sequence they were encountered during pre-training. This process helps us to assess the model's retention of information over time. We aim to assess if standard metrics like PPL can monitor forgetting trends throughout training by testing the model on this set.

3.1.1 Setup and PPL

We shuffled a dataset with 4.9e8 tokens subset from SlimPajama (Soboleva et al., 2023) for consistency across experiments, conducting standard and memory-replay pre-training. A test set was created by sequentially segmenting the training data according to the training steps and uniformly sampling 1/100 of each segment, *reflecting the model's*

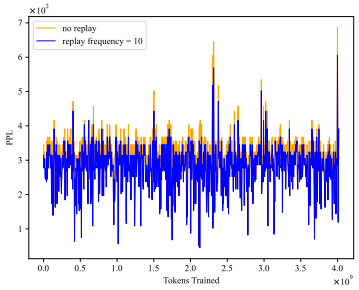


Figure 1: Perplexity (PPL) of the GPT-2 XL model on uniformly sampled 1/100 segments of the training data. Considering forgetting does help the performance.

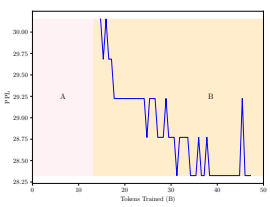
training progression. PPL is plotted against the number of training tokens processed, with the test set's token count scaled to match the model's exposure. More details are in Appendix B.1.

Results: The result is shown in Figure 1. Our observations indicate that: (1) The pre-trained model shows stable performance across early and late training data, with comparable perplexity (PPL), challenging the hypothesis of higher early training perplexity. This suggests either that forgetting is not occurring, contrary to our understanding, or that forgetting exists but is not captured by PPL. (2) Models with a replay mechanism during pre-training show better test set performance, with a notable drop in average PPL (280.66 with replay vs. 303.63 without), *indirectly confirming the existence of forgetting* through performance gains from repeated sample learning.

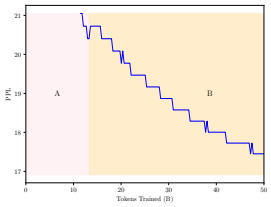
3.2 Underestimate of Pre-Training Forgetting

In previous experiments, we realized that detecting forgetting was challenging in a single pre-training dataset due to the *uniformity of the data*. To tackle this, we implement an A+B dual-dataset strategy, aiming for datasets A and B to be similar yet slightly different to magnify forgetting effects, aiding in metric assessment. With dataset A being much smaller than B, we aim to avoid overfitting on it. This emulates the scenario in an actual single pre-training dataset where a portion of the early data is at risk of being forgotten as the training progresses with a larger dataset. This is also a common and practical scenario for continuing pre-training.

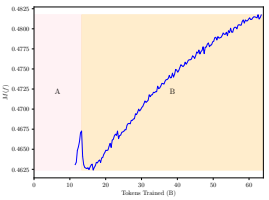
Setup: We proceed by uniformly sampling a subset from dataset A as a test set and then train on dataset B, evaluating the model to observe forgetting of dataset A. We conducted two experiments, employing the OpenWebText (Aaron Gokaslan*, 2019) dataset (∼8B tokens) in its entirety for dataset A in one experiment, and a uniformly sampled subset from the Pile (Gao et al., 2020) (∼ 13B) for the other. Dataset B was constituted by a uniformly sampled subset (∼ 49 B) tokens from SlimPajama. More details are in the Appendix B.2.



(a) PPL on OpenWebText



(b) PPL on the Pile



(c) M(f) on the Pile

Figure 2: (a), (b): Perplexity (PPL) of the eval of dataset A in relation to the number of trained tokens. B is a subset from SlimPajama. A is a subset of OpenWebText(a) or the Pile(b). The fluctuating PPL is not a good indicator of pre-training forgetting. (c): M(f) of the eval for the Pile. At the A-to-B dataset transition, M(f) shows negligible changes, and then M(f) consistently increases, where we capture the subtle signal of pre-training forgetting.

Results of PPL: The results in Figure 2 (a)(b) reveal an unexpected trend: contrary to expectations of increasing PPL for dataset A as a sign of forgetting during dataset B's training, the PPL for dataset A actually decreased in both setups. Even during the transition between datasets, PPL showed minimal signs of catastrophic forgetting.

3.2.1 M(f) Metric

Recognizing the shortcomings of perplexity in accurately measuring forgetting, we have turned to the M(f) metric introduced by Tirumala et al. (2022) for evaluation. The detailed definition of M(f) is:

Definition 1 *(Tirumala et al., 2022) Let V denote the vocabulary size. The set C consists of contexts (s, y), where s is an incomplete text and y is the correct token index. S contains all input contexts, and $f : S \to \mathbb{R}^V$ is a language model. A context c is memorized if $f(s)$'s maximum value corresponds to y, i.e., $\text{argmax}_{\mathbf{w}\in\mathbb{R}^V} f(s) = y$. We assess the fraction of contexts memorized by the model f using the metric $M(f) = \frac{\sum_{(s,y)\in C} \mathbb{1}\{\text{argmax}(f(s))=y\}}{|C|}$.*

Results of M(f): In this experiment, we continued to employ the A (the Pile) + B (SlimPajama) dataset setup and evaluated the model throughout the entire training process. We also continue to use a uniformly sampled 1/1000 part of A as the eval set. We observed that at the transition from dataset A to dataset B, M(f) exhibited negligible fluctuations. Subsequently, as training progressed on dataset B, the evaluation set's performance, as measured by M(f), demonstrated a continuous improvement. The results are given in Figure 2.

It is plausible to hypothesize that PPL's probabilistic averaging inherent may not accurately reflect forgetting for common tokens due to their high prediction accuracy, potentially masking information loss for less frequent elements. In contrast, the M(f) metric's binary evaluation is more sensitive to memory errors, offering a clearer view of the model's retention of critical information, essential for understanding catastrophic forgetting.

3.2.2 Limitation Leads to Underestimate

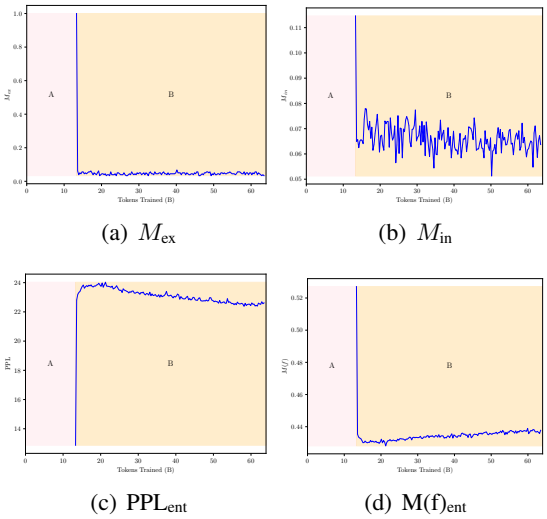
Certainly, it is important to acknowledge that both the PPL and M(f) metrics have limitations in fully capturing the model's forgetting behavior. Our observations indicate that throughout the training process, after the model has completed training on dataset A and transitions to dataset B, both metrics show a continuous improvement, with minimal signs of forgetting at the transition point. This suggests a plausible hypothesis: **The metrics' inability to account for the variability in data and token difficulty lead to an underestimation of forgetting, as they are dominated by features that are inherently resistant to forgetting**, such as common tokens and simple, everyday text. These features may not exhibit significant prediction errors when the dataset changes, thereby obscuring the true extent of the model's forgetting.

> **Takeaway 1:** PPL and M(f) metrics potentially mask true forgetting, as their bias towards easy-to-remember elements can underestimate the model's memory decline across dataset shifts.

4 New Entity-related Metrics for Measuring Pre-training Forgetting

4.1 The Essence of Pre-training Forgetting

Building upon the findings presented, a pertinent inquiry emerges: Which segments of the dataset should be scrutinized to gain a comprehensive understanding of the forgetting phenomenon?



(a) M_{ex} (b) M_{in}

(c) PPL_{ent} (d) M(f)_{ent}

Figure 3: Training dynamics across setting A (Pile) → B (SlimPajama) datasets: entity-focused evaluation set from A reveals marked metric degradation during the A-to-B transition. Despite this, traditional metrics on entity-focused samples such as PPL_{ent} and M(f)_{ent} exhibit partial recovery during dataset B training. This implies that even for entity-related evaluations, conventional metrics still largely focus on information that is less related to entities, which can continue to improve with further learning. Therefore, PPL_{ent} and M(f)_{ent} are not that sensitive and accurate as M_{ex} and M_{in} in measuring pre-training forgetting.

We argue that during pre-training, the focus should be on the forgetting associated with **entity-related information**. We posit that the capabilities imparted to a model by a dataset can be broadly categorized into two components: information related to entities and task-specific competencies. (1) As demonstrated by Sorscher et al. (2022), the power law scaling of error shows that many training examples are redundant, and in data-rich scenarios, pruning should focus on retaining challenging examples. Entity-related information, which is less frequent (Penedo et al., 2023), is crucial for users' perception of forgetting in LLMs, as it's harder to determine if the loss of abstract capabilities is due to model limitations or forgetting, making entity information key in pre-training. (2) We have also considered the approach of Supervised Fine-Tuning (SFT), which involves training pre-trained models on instructional data. This phase of training enhances the model's capabilities for downstream tasks, and we view it as a stage where the emphasis is on augmenting the model's competencies. Never-

theless, for the pre-training phase, our focus is more directed towards the acquisition of entity information. (3) Comparing with the forgetting of entities, the forgetting of other content, such as capabilities related to downstream tasks, is more challenging to define and remains ambiguous. Entities serve as an optimal vehicle for exploring the phenomenon of forgetting within our cognitive framework.

4.2 Our Proposed Entity-related Metrics

To evaluate the model's forgetting of entities, we follow the memorization score (Biderman et al., 2023a) and introduce additional metrics for pre-training forgetting. These additional metrics resemble entity-focused question answering. For further elaboration on the design and illustrative examples of our metrics, please refer to Appendix B.3.

(1) $\mathbf{M_{in}}$: Intuitively, this evaluates the model's capacity to recall entity-related to an entity details given its context. We select all samples S containing a set of entities C. For each sample $\mathbf{s_i} \in S$, we locate the entities and use the 32 preceding tokens as input, ensuring the entity $\mathbf{c_j} \in C$ is at the end. With this input $\mathbf{s_i}$, we then greedily decode 32 tokens $\hat{\mathbf{o}} = (o_1, o_2, ..., o_{32})$. Following this, we consider the next 32 tokens $(t_1, t_2, ..., t_{32})$ as our target output. We calculate the accuracy of the tokens, defined as $M_{\text{in}} = \frac{\sum_{\mathbf{s_j} \in S} \sum_{i=1}^{32} \mathbb{1}\{o_i = t_i\}}{32|S|}$.

(2) $\mathbf{M_{ex}}$: Intuitively, this metric tests if the model can recall an entity from the context where the entity is implied but not directly mentioned. Similar to M_{in}, for each sample $\mathbf{s_i}$ containing entity $\mathbf{c_j}$, we use the preceding 32 tokens as input (excluding $\mathbf{c_j}$) and the following 32 tokens as target output (starting with $\mathbf{c_j}$). After greedy decoding of 32 tokens $\hat{\mathbf{o}}$, we calculate $M_{\text{ex}} = \frac{\sum_{\mathbf{s_i} \in S} \text{is\_substring}(\mathbf{c_j}, \hat{\mathbf{o}})}{|S|}$, where $\text{is\_substring}(\mathbf{a}_1, \mathbf{a}_2)$ returns 1 if $\mathbf{a}_1$ is a substring of $\mathbf{a}_2$ and 0 otherwise.

Besides, we also adopt two entity-centric metrics $\mathbf{PPL_{ent}}$ and $\mathbf{M(f)_{ent}}$, which measure existing metrics PPL and M(f) on entity-involved samples.

Setup: In this section, we continue to leverage the A+B dataset configuration to accentuate the phenomenon of forgetting, employing the A (the Pile) + B (SlimPajama) dataset setup and training the model on both datasets. Testing is conducted during the training of dataset B.

We focused on entity-level forgetting by analyzing entity frequencies in datasets A and B, identifying a set of entities more frequently found in A. Using this set, we curated an evaluation set from A and monitored its metrics during B's training to measure the forgetting effect due to less exposure in B. See Appendix B.3 for details on the experiments.

Results: In Figure 3, we have demonstrated the following: (1) When evaluating forgetting on entity-related data, a significantly more pronounced decline is noted, with a notably slow recovery of metrics even during continued training. (2) In evaluations focusing on a subset of data that is rich in samples from source A compared to B, traditional metrics like PPL and M(f) suggest a recovery that may not fully capture the essence of forgetting. This apparent recovery may be due to less forgettable elements in the data. (3) Comparatively, the newly proposed metrics M_{ex} and M_{in} exhibit a more difficult recovery, which aligns closely with our expected manifestation of forgetting. This makes them more suitable for indicating forgetting.

> **Takeaway 2:** Our newly proposed entity-related metrics, M_{ex} and M_{in}, exhibit a more noticeable decline and difficult rebound, offering a clearer reflection of the forgetting phenomenon.

5 Memory Replay: A Simple Method for Alleviating Pre-training Forgetting

Inspired by the work of de Masson D'Autume et al. (2019), we introduce novel methods for episodic memory replay. We incorporate a module that retains a record of examples from the pre-training phase. During the learning period, we periodically draw a uniform sample from the memory's stored examples to conduct gradient updates.

5.1 Key Factors in Memory Replay

We have considered several potential design dimensions within the replay process, including the following:

- **Replay Frequency.** Following de Masson D'Autume et al. (2019), we match the size of our retrieved memory batches to our training batches. Given the computational intensity of replay, we execute a retrieval and gradient update every 100 steps, achieving an efficient 1% replay rate.
- **What to Store into Memory.** We consider strategies for memory sample storage: (1) including all samples encountered during pre-training, (2) prioritizing samples with entities for forgetting analysis,

and (3) choosing high-loss samples that may be more susceptible to forgetting. Advanced selection methods are reserved for future research.

- **Retrieve Strategy.** We've introduced two basic but impactful retrieval methods: random sampling and similarity-based sampling. Unlike de Masson D'Autume et al. (2019), who used a pre-trained BERT (Devlin et al., 2018) model for the latter, we opted for BM25 (Robertson et al., 2009), following its efficiency shown in TLM (Yao et al., 2022).
- **Exit Mechanism.** Given the fixed intervals of memory replay, the number of replayable samples is inherently limited. Simple replay strategies may lead to an imbalance in the samples being replayed, such as every replay batch coincidentally focusing on a few samples within the memory. Thus, we've implemented an exit mechanism to limit replay times, excluding them from further learning once they reach a set replay threshold.

5.2 Experimental Settings

In the previous section, we used two datasets, A and B, to study the forgetting effect. Now, to mimic a realistic pre-training setup, we've mixed and shuffled A with B into one complete set. We trained GPT2 from scratch using this combined set. To measure forgetting across the dataset, we took 1/5 of A+B, selected samples with entities, and made an evaluation set($\sim$ 200,000 samples). We then use the aforementioned 4 metrics to assess the results.

Although the ability to relearn past samples is beneficial, a significant drawback of the replay method is its increased training cost. Considering computational constraints and the need for simplicity, we have selected the following straightforward strategies, while leaving more sophisticated replay methods for future work:

- **Vanilla pre-training**. We use standard pre-training as a baseline.
- **Upper Bound**.We train from the vanilla pre-training checkpoint on the eval set, evaluating immediately to determine the model's peak memory retention.
- **BM25**. We leverage Elasticsearch (Elasticsearch, 2018) to maintain a memory of all encountered samples. At designated replay intervals, we match the current batch with stored samples based on similarity for retrieval, subsequently employing this data for replay.
- **BM25 + Samples with entities only**. During learning, we evaluate each sample for the presence of entities and only keep those in our memory for replay.
- **Focused Stochasticity: Constrained Entity Sampling with Exit Limit**. In this experiment, we shift from similarity-based retrieval to random sampling of previously learned samples at regular intervals. To prevent overlearning, we monitor replay frequency, excluding samples after they have been replayed 5 times.
- **Intensive Focused Stochasticity**: This variant of Focused Stochasticity intensifies the replay process, subjecting replayed samples to multiple epochs of learning. Further details on this method are elaborated in Section 6.2.2.

Method	PPL_{ent}	$M(f)_{ent}$	M_{ex} ($\times 10^{-3}$)	M_{in} ($\times 10^{-2}$)
Vanilla pre-training	26.03	0.4093	5.273	3.988
Upper Bound	23.74	0.4182	14.46	4.162
BM25	27.95	0.4015	4.586	3.895
BM25 + Samples with entities only	28.09	0.4013	4.575	3.941
Focused Stochasticity	25.79	0.4101	**5.496**	3.980
Intensive Focused Stochasticity	**25.40**	**0.4121**	5.450	**4.003**

Table 1: Evaluation results for replay strategies.

5.3 Effectiveness of Memory Replay

We display the evaluation in Table 1. The data indicates that similarity-based replay methods do not outperform the baseline, no matter if all samples or only those related to entities are kept in memory. We think this could be because these methods don't spread replay evenly; replaying all samples might focus too much on non-entity ones, while focusing only on entity-related samples could lead to too much attention on a specific subset, exaggerating the forgetting of other samples.

On the other hand, a simple sampling method improves upon the baseline, hinting that replay helps reduce forgetting during pre-training. However, there's still a big gap between the replay methods and the best possible performance, which means there's a lot of room to improve how we handle forgetting in pre-training.

> **Takeaway 3:** Our memory replay methods have shown potential in alleviating forgetting in the pre-training phase, while a gap persists relative to the upper bound, signifying the necessity for further research.

6 Explorations on Forgetting Curves

Observing the limitations of replay methods like Focused Stochasticity in the last section, we are led to explore opportunities for enhancing their efficacy. This exploration is motivated by the

renowned forgetting curve from human psychology (Loftus, 1985), which underscores the link between the intensity of learning and the pace of forgetting.

Recognizing that current methods involve samples being learned uniformly and at equal intervals with low intensity, we question the impact of transitioning to a strategy of intensive learning that focuses on specific information. We aim to explore the effect of short-term, high-frequency learning on the forgetting curve of large models, pondering whether models follow the same patterns as humans—where increasing the frequency of review slows down forgetting. With this understanding, we anticipate guiding the research on replay methods to enhance memory retention in models.

6.1 Setup

Exploring the nuances of memory retention and forgetting in LLMs, we focus on two critical inquiries: (1) **Learning intensity's impact**: We explore the hypothesis that increased initial learning intensity may result in more robust memory retention, potentially flattening the forgetting curve. (2) **Memorability and memory durability**: We determine if challenging samples, post-intensive learning, remain at risk of forgetting during pre-training.

6.2 Results on LLMs' Forgetting Curves

To tackle these inquiries, we first select samples related to entities of interest. After the model undergoes an initial epoch of pre-training, we subject these samples to intensive training across several epochs. The purpose of the initial pre-training epoch is to ensure the model reaches a baseline level of language proficiency. This step is crucial to prevent general language ability improvements from confounding the experiment, allowing for a clear focus on the forgetting phenomenon rather than overall model enhancement.

Post the intensive learning phase, these entity-related samples serve as our evaluation set. As we proceed with pre-training, we continuously assess this set using our established metrics to monitor the forgetting curve. This ongoing evaluation allows us to track how the memory of these samples evolves and to understand the interplay between initial learning intensity and long-term retention within the context of LLM pre-training. For further details on this experimental design, please refer to the Appendix B.4.

6.2.1 Initial Learning Intensity and Forgetting Curves

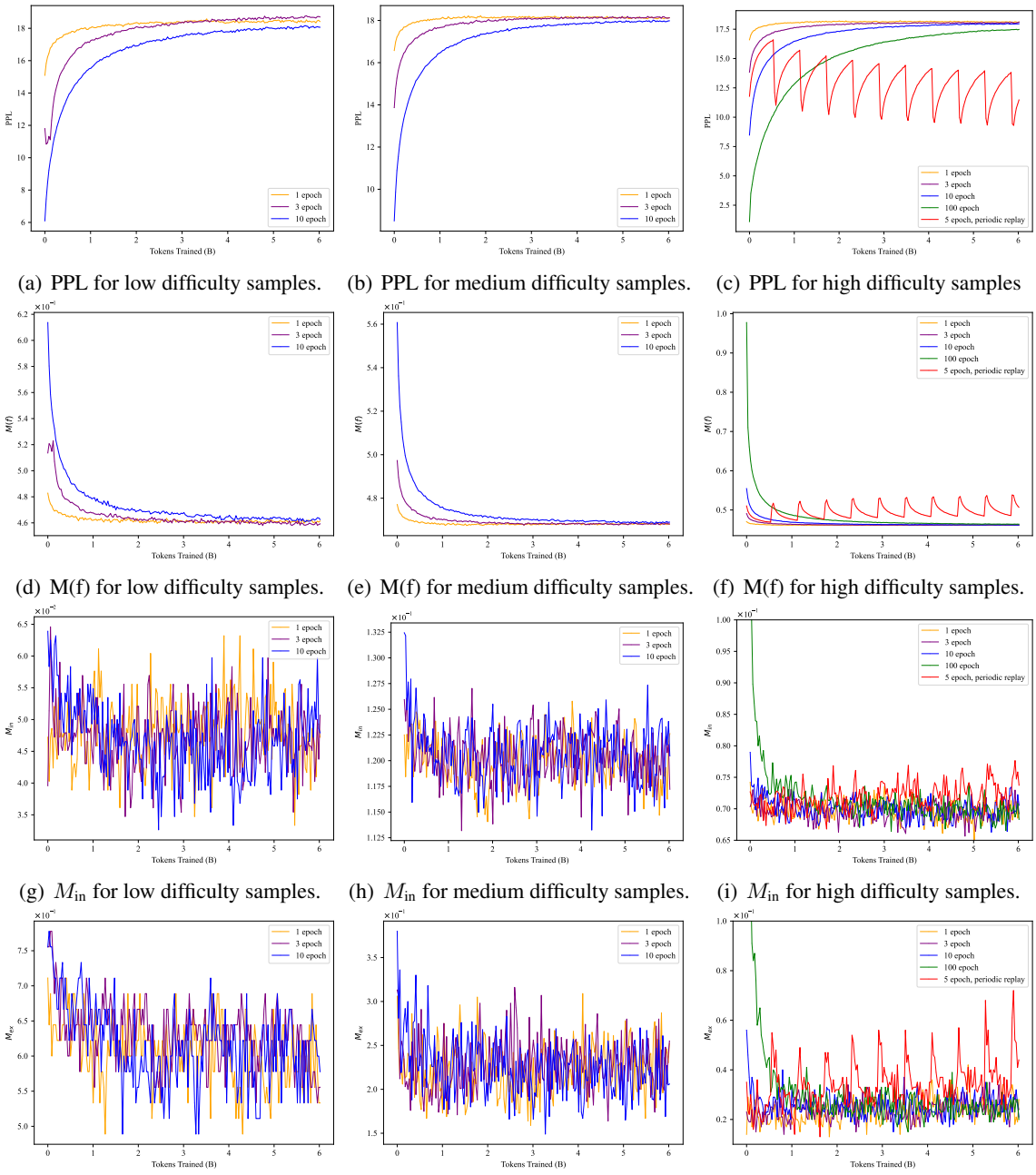
As shown in Figure 4, our experiments indicate that higher initial learning intensity results in better performance across various metrics, yet as further pre-training occurs, the results from experiments with lower initial learning intensity tend to catch up. This pattern mirrors human learning curves (Loftus, 1985), and we offer a detailed comparison in Appendix C. Over the learning period, a divergence is observed; experiments with a very high initial learning intensity maintain a gap compared to those with a lower initial intensity. This gap is more pronounced for less difficult data, while for more challenging data, the effects of learning even out. This suggests that data that are more difficult to memorize benefit from more intensive learning to achieve enhanced memory retention.

6.2.2 Periodic Intensive Replay

Building on our findings and the human ability to reduce forgetting through periodic, intense learning, we aim to (1) assess the impact of periodic, intensive replay on a model's forgetting curve, and (2) determine if this can enhance aforementioned memory replay methods. To delve deeper into these effects, we have focused our experiments on the most challenging samples. After the initial phase of high-intensity learning, we have introduced a replay process in the ongoing pre-training. This process involves revisiting the samples every 1000 steps, with each replay session consisting of 5 epochs of learning.

In this experiment, the replay intervals were relatively large, which was acceptable in terms of efficiency. Moreover, the replay method outperformed the baseline. Although there was a temporary decline after each replay, the overall performance saw improvement over time. We discovered that periodic, high-intensity replay on the forgetting curve leads to an enhancement of both the upper and lower bounds. Moreover, this approach proved more effective and cost-efficient than directly replay with 100 epochs.

Thus, we believe that such human-like strategies could guide the design of replay mechanisms in pre-training. To test this hypothesis, we conducted an experiment and enhanced the Focused Stochasticity method in Section 5.2. Specifically, we intensified the learning process for each replay batch, with each batch undergoing five epochs of learning. The approach, denoted as Intensive Focused



(a) PPL for low difficulty samples. (b) PPL for medium difficulty samples. (c) PPL for high difficulty samples

(d) M(f) for low difficulty samples. (e) M(f) for medium difficulty samples. (f) M(f) for high difficulty samples.

(g) M_{in} for low difficulty samples. (h) M_{in} for medium difficulty samples. (i) M_{in} for high difficulty samples.

(j) M_{ex} for low difficulty samples. (k) M_{ex} for medium difficulty samples. (l) M_{ex} for high difficulty samples.

Figure 4: Metrics of samples categorized by difficulty level, with columns representing distinct difficulty levels and rows indicating different initial learning intensities.

Stochasticity in Table 1, surpasses the baseline in every metric, indicating its potential in reducing pre-training forgetting.

Takeaway 4: The parallels in forgetting patterns between humans and LLMs suggest that periodic, intensive replay could be key to mitigating memory loss. Experiments conducted during the pre-training phase have also confirmed this point.

7 Conclusion and Future Work

Our research sheds new light on catastrophic forgetting in LLMs during pre-training. We scrutinized traditional metrics, introduced novel ones for a clearer analysis of forgetting, and proposed memory-replay techniques to bolster knowledge retention. Additionally, we explored the forgetting curve post-intense, short-term learning, uncovering similarities with human memory decay, offering insights into information retention dynamics.

8 Limitations

Our investigation into catastrophic forgetting within the pre-training phase of LLMs, while pioneering, is bounded by computational limitations. The experimental requirements, estimated at approximately 10,000 GPU hours for execution on 8 NVIDIA A100 GPUs with 40 GiB VRAM, present a significant challenge. This constraint inevitably limits the scale of our experiments, making it challenging to verify with larger models and different datasets.

Informed by the scaling law (Kaplan et al., 2020), we recognize that our findings from a smaller model may provide valuable insights for larger-scale experiments. This framework indicates that our study could contribute to the design of future research, acknowledging the limitations in scaling our results.

Our approach to memory replay has shown potential in alleviating catastrophic forgetting, but there is still room for improvement in terms of its effectiveness. Our investigation did not delve deeply into the granular effects of each variable on the experimental outcomes. The complexity of memory replay mechanisms requires a more nuanced analysis to fully understand how different factors interplay and influence the results.

Additionally, the concentrated learning of memory replay, while beneficial, may engender trade-offs that affect the model's generalizability. We hypothesize that the focused emphasis of certain data subsets could lead to a diminished capacity for the model to adapt to tasks beyond the focused areas, such as numerical data processing or other cognitively distinct downstream tasks.

We recognize that forgetting in pre-training differs from that in SFT, each requiring distinct metrics and methods for mitigation. Yet, there are connections between them. In future work, we also aim to explore the impact of our methods on forgetting in downstream tasks.

Despite these limitations, our study exemplifies the scientific endeavor to confront complex problems with rigor and without reservation. Our work is a courageous step towards understanding the intricate processes of memory retention and forgetting in LLMs, reflecting a sincere commitment to advancing our collective knowledge, even in the face of substantial challenges.

References

Ellie Pavlick Stefanie Tellex Aaron Gokaslan*, Vanya Cohen*. 2019. Openwebtext corpus.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raf. 2023a. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023b. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*. PMLR.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. 2022. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security*.

C Samuel Craig, Brian Sternthal, and Karen Olshan. 1972. The effect of overlearning on retention. *Journal of General Psychology*.

Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. *NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

BV Elasticsearch. 2018. Elasticsearch. *software], version*.

Wikimedia Foundation. Wikimedia downloads.

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

Kshitij Gupta, Benjamin Thérien, Adam Ibrahim, Mats L Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re) warm your model? *arXiv preprint arXiv:2308.04014*.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2021. Lifelong pretraining: Continually adapting language models to emerging corpora. *arXiv preprint arXiv:2110.08534*.

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. *arXiv preprint arXiv:2001.08361*.

Geoffrey R Loftus. 1985. Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*. Elsevier.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Yujia Qin, Jiajie Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. Elle: Efficient lifelong pre-training for emerging data. *arXiv preprint arXiv:2203.06311*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE.

Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *NeurIPS*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *AAAI*.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *NeurIPS*.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023b. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*.

Zirui Wang, Sanket Vaibhav Mehta, Barnabás Póczos, and Jaime Carbonell. 2020. Efficient meta lifelong-learning with limited memory. *arXiv preprint arXiv:2010.02500*.

Chengyue Wu, Yukang Gan, Yixiao Ge, Zeyu Lu, Jiahao Wang, Ye Feng, Ping Luo, and Ying Shan. 2024. Llama pro: Progressive llama with block expansion. *arXiv preprint arXiv:2401.02415*.

Yinjun Wu, Edgar Dobriban, and Susan Davidson. 2020. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*. PMLR.

Xingcheng Yao, Yanan Zheng, Xiaocong Yang, and Zhilin Yang. 2022. Nlp from scratch without large-scale pretraining: A simple and efficient framework. PMLR.

A Further Discussions on Pre-training Forgetting

In this section, we discuss the intuition and methodology behind the paper, as well as potential issues.

1. **Why should we be concerned about model forgetting at the sample level during pre-training?**
Developers and researchers have frequently observed that large models, despite their extensive deployment, are prone to errors in factual domains, especially concerning entity-related information (Wang et al., 2023a). These discrepancies can substantially affect user perception and trust. However, there is a scarcity of research on the influence of learning during the pre-training phase on this type of information, and even less on how models remember and forget information during pre-training. The phenomenon of sample-level forgetting in pre-training is also difficult to define clearly, analyze, and further explore.
2. **How should we understand entity-related metrics, and why is it important to focus on forgetting at the entity level?**
(1) Forgetting across the entire pre-training dataset is extremely difficult to define and study, hence we concentrate on a specific subset. Errors related to entity information are easily noticeable in model applications and significantly impact user experience. (2) Beyond the model's memory of entity information, we also consider its capabilities during pre-training, especially since the Supervised Fine-Tuning (SFT) phase places more emphasis on instructional data. This phase enhances the model's competencies for downstream tasks, and we see it as a stage for augmenting the model's capabilities. Therefore, we believe the pre-training phase should place greater emphasis on exploring entity information. (3) In Section 3.2, we demonstrate that overall data forgetting is hard to evaluate, as there is no clear decline in model performance when switching training data (we deliberately selected parts of data from A to ensure minimal repetition in B), and almost no change in metrics is observed during the switch. Instead, during training in B, the model's capabilities continue to improve, even surpassing the metrics achieved during training in A, which contradicts the intuition of forgetting. PPL does not intuitively reflect the model's forgetting; in contrast, the metrics concentrated on entities show significant changes on entity-related data, with almost no recovery, facilitating the direct study of the forgetting phenomenon.
3. **Since the model may leak verbatim sequences of personal information, is sample-level forgetting harmful?**
In our study, we focus on learning and the retention of factual information related to entities, which models should not forget and that is prevalent in the pre-training data. We diverge from concerns about leaking verbatim personal information. There is extensive literature on machine unlearning (Wu et al., 2020; Bourtoule et al., 2021; Chen et al., 2022), which typically addresses scenarios involving privacy protection and changes in user information. These scenarios fall outside the scope of our work, although our research might offer insights into the design of machine unlearning methods.
4. **Is this study primarily addressing hallucinations, or is it actually more focused on the model's tendency to forget entity-related information rather than producing false outputs?**
Our research concentrates on the model's inclination to forget information pertaining to entities, diverging from the generation of erroneous outputs, commonly known as hallucinations. However, it is true that our work offers a perspective on the concept of hallucinations, where the two newly designed metrics, M_{ex} and M_{in}, can be interpreted as potential false negatives and false positives in the pre-training model's responses: the model, given relevant information, fails to identify the correct entity; or the model provides an entity and some information but is unable to supply the related context.
5. **Should we expect an LLM to reproduce exact training text, given it's not a lossless compression model?**
In our study, we do not anticipate LLMs to reproduce the exact training text. Specifically, our M_{ex} metric solely assesses whether the ground truth entity is included in the output; while capturing the formalization of information related to the entity presents challenges. For the M_{in} metric, we follow the design of Biderman et al. (2023a), calculating accuracy for each token. We consider that alternative design schemes might be possible, such as utilizing a BERT model (Devlin et al., 2018) to calculate the similarity between the generated tokens and the ground truth tokens. We have reserved this exploration for future research.

B Setup Details

In this section, we outline our experimental setup. We selected a batch size of 576, informed by our use of 8 NVIDIA A100 GPUs with 40 GiB VRAM, and aligned with GPT-2‘s (Radford et al., 2019) hyperparameter recommendations for optimal performance on our hardware configuration. A consistent sequence length of 1024 was applied across all experiments. Training is executed in half-precision format using BF16, and we capitalize on the Zero Redundancy Optimizer (ZeRO) Stage 2 (Rajbhandari et al., 2020) to enable efficient scaling across multiple machines. We draw inspiration from the works of Biderman et al. (2023b); Gupta et al. (2023); Radford et al. (2019), employing a cosine learning rate decay that reduces to a minimum of 0.1 times the Maximum Learning Rate (MaxLr), with the MaxLr itself set at 6×10^{-4}.

B.1 Setup for Section 3.1

We utilized the GPT-2 XL model (1.5B) (Radford et al., 2019) and trained it on a dataset sampled from SlimPajama (Soboleva et al., 2023), consisting of 4.9e8 tokens. Prior to training, we shuffled the data to ensure that the order of training instances was consistent across different experiments. We conducted two experiments: a standard pre-training and a pre-training with a replay mechanism that retrieves a batch of data, equivalent in size to the training batch. (where we stored all trained data using Elasticsearch (Elasticsearch, 2018) and performed a replay every 10 steps). At each replay step, we use the current batch‘s training data to uniformly sample an equal amount of data from the completed training data based on similarity. This ensures a uniform replay throughout the entire data training process, with an additional 1/10 increase in computational budget. For evaluation, we constructed a test set by sequentially segmenting the training data according to the training steps and uniformly sampling 1/100 of each segment. The samples were then reassembled in their original stepwise order to ensure uniform distribution across the training steps, thus creating a test set that mirrors the model‘s training progression. We plotted perplexity (PPL) against the number of training tokens processed, with the evaluation set‘s token count scaled proportionally to reflect the model‘s exposure to the training data.

B.2 Setup for Section 3.2

To ensure computational feasibility in our experiments, we choose GPT-2 (0.1B) in this section. We uniformly sample 1/1000 of dataset A to constitute a eval set, and perform evaluations every 1000 training steps during the training process of dataset B.

B.3 Setup for Section 4.2

Sampled entities
‘ Terrel Bell‘, ‘ BIST‘, ‘ The Great Hunt‘, ‘ Best in Drag Show‘, ‘ Stella Maris‘, ‘ William Knighton‘, ‘ Italian campaign‘, ‘ The Octopus Project‘, ‘ Light Cycle‘, ‘ Clark Street‘, ‘ Paulette Hamilton‘, ‘ Robert Mack‘, ‘ Nusrat‘, ‘ Soul Catcher‘, ‘ Lord of Light‘, ‘ Bieger‘, ‘ Foreach loop‘, ‘ Choruss‘, ‘ Screen space ambient occlusion‘, ‘ Florida Department of Environmental Protection‘, ‘ USA Ultimate‘, ‘ Historical Association‘, ‘ Robert Holt‘, ‘ Willie Nile‘, ‘ Fiordland National Park‘, ‘ Star Wars: The Clone Wars‘, ‘ Crouch End‘, ‘ Tracy Ham‘, ‘ Jimmy Chamberlin‘, ‘ Journal of Food Science‘, ‘ Comet Tempel‘, ‘ AirMed International‘, ‘ CanWaCH‘, ‘ Pumapunku‘, ‘ Pre-law‘, ‘ Arovane‘, ‘ Diex‘, ‘ Her Escape‘, ‘ Voltige‘, ‘ Triadelphia‘, ‘ Florian Zeller‘, ‘ The Busy World of Richard Scarry‘, ‘ Texting while driving‘, ‘ Amir Wilson‘, ‘ Julie White‘, ‘ Lenox‘, ‘ GNPDA2‘, ‘ Cammie Dunaway‘, ‘ Session Man‘, ‘ Charoen Krung Road‘, ‘ James Raine‘, ‘ Archie Andrews‘, ‘ The Picture of Dorian Gray‘, ‘ Theresa Caputo‘, ‘ Schauinslandbahn‘, ‘ Japanese relocation‘, ‘ O.C. Handa‘, ‘ Afula‘, ‘ The Secrets‘, ‘ Sonnet 61‘, ‘ Daniel Bell‘, ‘ The Dawn‘, ‘ Bob Berry‘, ‘ Bigger Life‘, ‘ Jamaica Wine House‘, ‘ Conica‘, ‘ Renuar‘, ‘ Plantation, Florida‘, ‘ Fasser‘, ‘ Al-Qadi‘, ‘ Vassy‘, ‘ Tom Dempsey‘, ‘ Department of Agriculture, Environment and Rural Affairs‘, ‘ Abdallah Djaballah‘, ‘ Silent Hill 2‘, ‘ Bill Ayres‘, ‘ Jeremy Howe‘, ‘ J15‘, ‘ Jake Ryan‘, ‘ Black Mafia‘, ‘ Nicholas Fox‘, ‘ Interstate 78‘, ‘ Mark Stein‘, ‘ Pietro Torri‘, ‘ Wet sump‘, ‘ Centre national des arts plastiques‘, ‘ Nitro Express‘, ‘ Wyvill‘, ‘ WSRA‘, ‘ Whitewater River‘, ‘ Merry Christmas Mr. Lawrence‘, ‘ Jon Jansen‘, ‘ Le Message‘, ‘ Mavrommati‘, ‘ Tourouvre‘, ‘ Bob Peterson‘, ‘ America Again‘, ‘ Livernois‘, ‘ The Shepherd Express‘, ‘ Hypercalcaemia‘

Table 2: Sampled entities from English Wikipedia.

We followed Biderman et al. (2023a), selecting a sequence length of 32 for both the input and output

Prompt	True Continuation	Greedily Generated Sequence	M_{in}
The Amazon Rainforest ,	known as the Earth's lungs	known as the Moon's lungs	$\frac{1+1+1+0+1}{5} = 0.8$
The Amazon Rainforest ,	known as the Earth's lungs	known as the Moon's legs	$\frac{1+1+1+0+1}{5} = 0.6$
The Colosseum in Rome, also known as the Flavian Amphitheatre ,	is an iconic symbol of the Roman Empire's architectural prowess.	is an iconic symbol of the Russian Federation's scientific prowess .	$\frac{1+1+1+1+1+1+0+0+0+1}{10} = 0.7$

Table 3: Examples of M_{in} calculation with different prompts. These samples are provided for illustrative purposes and are not from the real training data.

Entity	Prompt	True Continuation	Greedily Generated Sequence	M_{ex}
Leonardo da Vinci	The Mona Lisa, painted by	Leonardo da Vinci , is renowned for its elusive	Leonardo da Vinci , is renowned for its elusive	1
Leonardo da Vinci	The Mona Lisa, painted by	Leonardo da Vinci , is renowned for its elusive	a man called Leonardo da Vinci , is renowned for	1
Leonardo da Vinci	The Mona Lisa, painted by	Leonardo da Vinci , is renowned for its elusive	Donald Trump , is renowned for its elusive	0
the United States	The Statue of Liberty, a gift from France to	the United States , stands as a symbol	the world, mysteriously appeared on an uninhabited island	0
the United States	The Statue of Liberty, a gift from France to	the United States , stands as a symbol	tell the enduring friendship with the United States	1

Table 4: Examples of M_{ex} calculation with different prompts. These samples are provided for illustrative purposes and are not from the real training data.

of our M_{ex} and M_{in} metrics. We collected entities from English Wikipedia dataset (Foundation). Some randomly sampled entities are shown in Table 2.

To spotlight entity-level forgetting, we evenly sampled 400,000 English Wikipedia entries, comparing entity frequencies in datasets A and B. We selected the intersection C of entities that were top 1/2 frequent in A and bottom 1/2 in B to accentuate the distribution disparity. Samples from A with entities in C constituted our evaluation set. Following the approach of Biderman et al. (2023a), we retained a subset where $M_{\text{ex}} = 1$ post A's training to scrutinize their forgetting during B's training.

We provide illustrative examples in Table 3 and Table 4 to provide clearer explanations of M_{in} and M_{ex}.

B.4 Setup for Section 6.2

It is evident that M_{ex} assigns a binary label to each sample: a label of 1 is given if the ground truth entity appears within the generated 32 tokens, and a 0 is assigned otherwise. Utilizing the challenging metric of M_{ex}, we can categorize the difficulty of data memorization as follows: We performed an evaluation on the portion of the pre-training data that includes entities, recorded each entity alongside the samples that received labels of 1 or 0, and then calculated the accuracy rate for each entity based on these labels. We then divided the entities into groups with roughly equal accuracy rates, ensuring that during the phase of intensive, short-term learning, the related samples for certain entities are the focus of concentrated study. For the data categorized into different difficulty levels, we carried out experiments with varying degrees of learning intensity—specifically, by adjusting the number of epochs dedicated to this phase of learning.

C Comparison of Forgetting Curves between Humans and LLMs

The reproduced human forgetting curve, originally reported by Craig et al. (1972), is illustrated below, reflecting the typical decline in memory retention over time. In their study, 180 undergraduates participated in an experiment involving exposure to magazine advertisements under controlled conditions. They were categorized into three groups based on the extent of learning: 100%, 200%, and 300%, determined by the number of 5-second repetitions of 12 ads. Following exposure, 15 participants from each group were assigned to one of four retention tests occurring at immediate, 1-day, 7-day, or 28-day intervals. The study utilized a 3 $\times$ 4 factorial design, assessing the impact of learning intensity and retention intervals on the recall of brand names. It can be observed that there are similarities between the model's forgetting curve and the human forgetting curve, with higher initial learning intensity resulting in a relatively slower rate of forgetting.

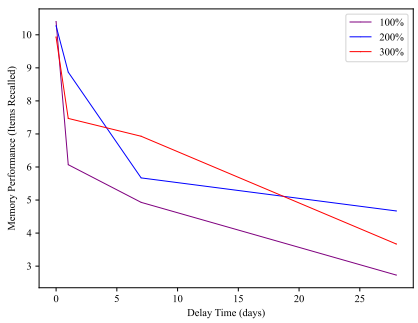


Figure 5: Human forgetting curve from Craig et al. (1972).