Exploring Forgetting in LLM Pre-Training

Anonymous ACL submission

Abstract

In large language model (LLM), 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 a dearth of studies focusing on finer-grained forgetting at the sample level during pre-training. This paper delves into the intricacies of forgetting during the pre-training phase, where models are typically trained on a massive diverse corpus for only one epoch. We systematically explore the existence, essence, and measurement of forgetting in LLM pre-training. Specifically, we investigate the limitations of traditional metrics such as perplexity (PPL) in accurately measuring forgetting in pre-training, and propose several new metrics evaluating LLM's ability to assess related memories of entities, which is viewed as the key reflection of whether forgetting happens in pre-training. Extensive evaluations and insights on forgetting of pre-training facilitate future research on LLMs.

1 Introduction

In NLP, the phenomenon of catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990) poses a significant challenge to the development of models capable of continuous learning, which is also observed in LLM. 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 can lead to the model's inability to remember past information effectively.

Although some pioneer efforts have explored the forgetting issue in LLM fine-tuning (which focuses more on task-level 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. (2023), 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., 2023) studied sample-level memorization, where some experiments roughly 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.

Systematically exploring the forgetting in LLM pre-training is essential, and it does widely exist in current LLMs, harming their performance. Intuitively, a typical situation when we notice there is certain forgetting happening in pre-training is that: LLM outputs an unsatisfactory reply, while the hint is already in the pre-training data. However, quantitatively measuring forgetting in pre-training is extremely difficult. Different from 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 like perplexity (PPL) are also verified to be insensitive in measuring forgetting in pre-training. This raises two pertinent questions: (1) *How to correctly recognize the forgetting in pre-training?* (2) *How to quantitatively measure it in pre-training?*

To address the above questions, we first build a specialized scenario to magnify the forgetting issue, 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 the most explicit and significant indicator to reflect pre-training forgetting for practical user perception. Subsequently, we propose three novel metrics and experimentally confirm the occurrence of forget-

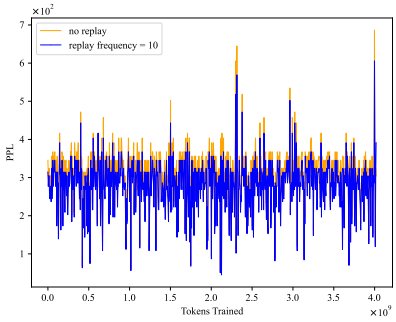


Figure 1: Perplexity (PPL) of the GPT-2 XL model on uniformly sampled 1/100 segments of the training data. Considering forgetting does help the performance.

ting during pre-training.

Our contributions are summarized as follows: (1) We systematically highlight the existence and essence of forgetting in pre-training. (2) We introduce several novel entity-related metrics to quantitatively measure pre-training forgetting.

2 Existence of Pre-training Forgetting

2.1 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. Consequently, a natural approach to testing this hypothesis is to sample the training data uniformly in the order of their presentation during pre-training to form an evaluation set. We aim to evaluate whether conventional metrics like PPL can effectively track the trend of forgetting over training steps by assessing the model's performance on this set.

Setup: 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 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.

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.

2.2 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 address the difficulty of detecting forgetting in a uniform single pre-training dataset, we've adopted an A+B dual-dataset approach. This setup, with dataset A's smaller subset and dataset B's larger subset, mimics the incremental addition of data in pre-training, magnifying forgetting effects for clearer metric evaluation. This is a common and practical scenario for continuing pre-training.

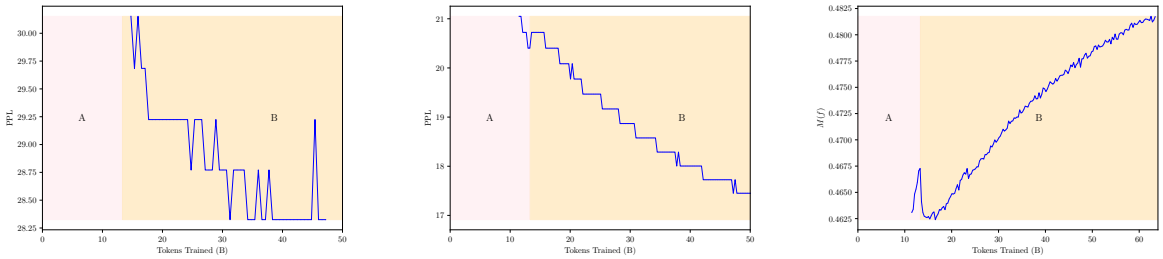
Setup: We proceed by uniformly sampling a subset from dataset A as a test set and then train on dataset B, evaluating the model to observe its response to the changed data distribution, offering insights into knowledge retention and decay.

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 of approximately 49 B tokens from SlimPajama. More details are in the Appendix.

Results of PPL: The results are shown in Figure 2 (a)(b). Contrary to our initial assumptions that the perplexity (PPL) of dataset A would gradually increase during the training of dataset B due to forgetting, our observations show that the PPL on the A evaluation set actually decreased progressively throughout the training of dataset B in both experimental setups. Even at the critical juncture when transitioning from dataset A to dataset B, there was a negligible indication of catastrophic forgetting detectable through PPL.

M(f) Metric: Acknowledging the limitations of perplexity as a metric for capturing forgetting, we have adopted the M(f) metric proposed by Tirumala et al. (2022). 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*



(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.

(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 test 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.

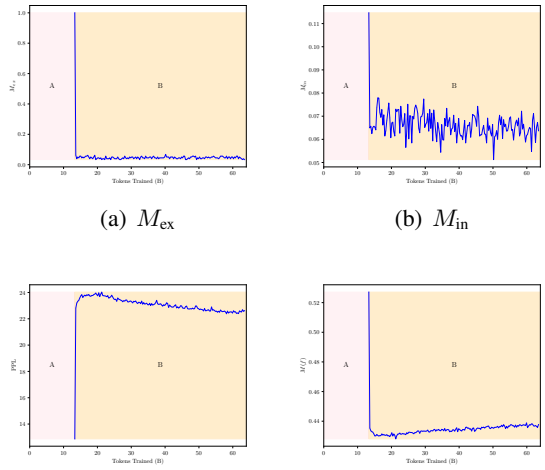
It is plausible to hypothesize that PPL's probabilistic averaging inherent may not accurately reflect forgetting for common tokens due to their high prediction accuracy, potentially masking information loss for less frequent elements. In contrast, the M(f) metric's binary evaluation is more sensitive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

Limitation leads to Underestimate: Certainly, it is important to acknowledge that both the perplexity (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 may 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). Such features may not exhibit significant prediction errors even when the dataset changes, thereby obscuring the true extent of the model's forgetting.

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

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?

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 suggests that many training examples are redundant, and in data-rich scenarios, pruning should focus on retaining challenging examples. Entity-related data, which is less frequent, 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.



(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.

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. Nevertheless, 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 (e.g., capabilities related to downstream tasks) by the model is more challenging to define and remains ambiguous. Entities serve as an optimal vehicle for exploring the phenomenon of forgetting within our cognitive framework.

Entity-related Metrics: To evaluate the model's forgetting of entities, we followed the memorization score in Biderman et al. (2023) and introduced two additional metrics for pre-training forgetting.

(1) $\mathbf{M_{in}}$: For a set of entities C, we select all samples S containing these entities, determine their positions in each sample $s_i \in S$, and use the preceding 32 tokens as input (with the entity $c_j \in C$ at the end) and the following 32 tokens as output. We then decode 32 tokens greedily and measure accuracy against the output, assessing the model's memory of entity-related information.

(2) $\mathbf{M_{ex}}$: Similar to M_{in}, for each sample s_i containing entity c_j, we use the preceding 32 tokens as input (excluding c_j) and the following 32 tokens as output (starting with c_j). Greedy decoding of 32 tokens yields $\hat{o}$, and we score 1 if $\hat{o}$ includes c_j, 0 otherwise, assessing the model's recall of entities given related context.

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.

To focus on entity-level forgetting, we selected 400,000 Wikipedia entries, analyzing entity frequency across datasets A and B. We formed set C from the top 1/2 frequent entities in A and the bottom 1/2 in B. Samples extracted from A with entities from C were used to evaluate dataset A. Due to M_{ex}'s complexity, we kept samples with $M_{\text{ex}} = 1$ post-training on A and monitored their forgetting during B's training.

Results: In this experiment, we have demonstrated the following: (1) When testing for forgetting on data related to entities, a more pronounced forgetting phenomenon is observed. (2) Regardless of whether perplexity (PPL) or M(f) is used, the metrics show a gradual recovery over the course of training, indicating that these metrics are more influenced by the less forgettable aspects of the data. (3) Comparatively, the newly proposed metrics M_{ex} and M_{in} are more challenging to recover, making them more suitable for indicating the phenomenon of forgetting.

4 Conclusion and Future Work

In conclusion, our research contributes to the understanding of catastrophic forgetting during the pre-training phase of large language models. By examining the limitations of traditional metrics and introducing new ones, we have provided a more detailed analysis of the forgetting phenomenon.

In the future, to mitigate the phenomenon of forgetting, it is necessary to investigate (1) the impact of more refined data ratios and learning sequences in pre-training datasets, and (2) the potential of memory-replay methods to alleviate forgetting.

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. 2023. Emergent and predictable memorization in large language models. *arXiv preprint arXiv:2304.11158*.

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

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

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*.

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*.

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.

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*.

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

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*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023. 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*.

A Setup Details

We include the detailed experimental setup in this section. For all experiments, we set the training micro-batch size as 576, and fix a sequence length of 1024 across all experiments.

Setup for Section 2.1 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 training volume. 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.

Setup for Section 2.2 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.