Demystifying Language Model Forgetting with Low-Rank Example Associations

Xisen Jin, Xiang Ren
University of Southern California
{xisenjin,xiangren}@usc.edu

Abstract

Large Language models (LLMs) suffer from forgetting of upstream data when fine-tuned. Despite efforts on mitigating forgetting, few have investigated whether, and how forgotten upstream examples are associated with newly learned tasks. Insights on such associations enable efficient and targeted mitigation of forgetting. In this paper, we empirically analyze forgetting that occurs in N upstream examples (of language modeling or instruction-tuning) after fine-tuning LLMs on one of M new tasks, and visualize their associations with a $M \times N$ matrix. We empirically show that the degree of forgetting can often be approximated by simple multiplicative effects of the upstream examples and newly learned tasks. We also reveal more complicated patterns where specific subsets of examples are forgotten. Following our analysis, we predict forgetting that happens on upstream examples when learning a new task with matrix completion over the empirical associations, outperforming prior approaches that rely on trainable LMs. Replaying predicted examples can statistically significantly improve over random examples for alleviating forgetting.

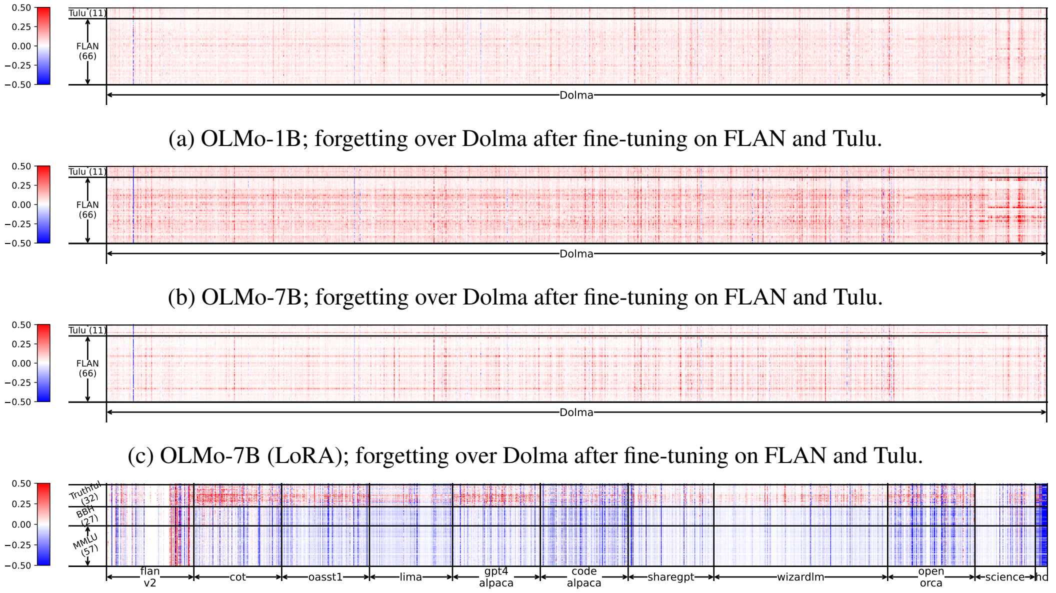
1 Introduction

There has been a growing need for long-term usability of LLMs. While fine-tuning allows incremental adaptation of models, it risks catastrophic forgetting [23] of upstream knowledge learned in the pre-training phase, causing unintended prediction changes over known information. This is problematic for stability of online deployed LLM systems, limiting the practical feasibility of continual fine-tuning.

Extensive works have studied algorithms to mitigate forgetting [33]. Some works analyze patterns of frequently forgotten examples [39, 22, 46] or effects of models and hyperparameters [25, 15, 27, 11]. However, not many have explored how the associations between learned tasks and upstream examples inform forgetting. Theoretical and empirical study reveals associations between learned and forgotten tasks in shallower neural networks [19, 6, 28], but such associations are under-explored in LLMs, or measured for upstream data of language modeling or instruction-tuning.

In this paper, we empirically study associations between learned tasks and forgotten upstream examples (of language modeling or instruction-tuning). We preform statistics of forgetting (in log perplexity increase) over N upstream examples, after fine-tuning the model on one of the M new tasks, represented in a $M \times N$ matrix. We experiment with OLMo-1B, OLMo-7B and OLMo-7B-Instruct [8] models where upstream data is released open-source. We fine-tune LLMs on diverse and unseen instruction-tuning tasks and measure forgetting on upstream data. Afterwards, we visualize the matrices and fit the observations with statistical models to analyze the associations.

Our findings suggests that approximating forgetting with a simple multiplicative scalar effects of learned tasks and upstream examples in LLMs results in decent R^2 between 0.49 and 0.76 depending on the models and data. More complicated associations are revealed through visualization



(a) OLMo-1B; forgetting over Dolma after fine-tuning on FLAN and Tulu.

(b) OLMo-7B; forgetting over Dolma after fine-tuning on FLAN and Tulu.

(c) OLMo-7B (LoRA); forgetting over Dolma after fine-tuning on FLAN and Tulu.

(d) OLMo-7B-Instruct (LoRA); forgetting over Tulu after few-shot fine-tuning on MMLU, BBH, and TruthfulQA.

Figure 1: Visualized matrices of assoications between learned tasks and forgotten examples. We plot forgetting (log-perplexity increase) that occurs on an upstream example (in x-axis) after learning a new task (in y-axis). Log-perplexity increase can be zero or negative, indicating no forgetting.

and statistics. We see the associations are more complicated in OLMo-7B than OLMo-1B under identical training configurations; the associations also become simpler as we perform LoRA [10] fine-tuning compared to full-parameter tuning. Following our analysis, we propose to predict example forgetting on unseen tasks as a matrix completion problem over the association matrices analogical to collaborative filtering [30] in recommender systems, achieving efficiency and interpretability. Our k-nearest neighbor (KNN) model outperforms prior approaches that learns semantic relations of two examples with LMs [14] . We verify the benefit of prediction by upweighting examples with higher predicted forgetting during replay as we fine-tune LLMs on new instruction-tuning tasks, achieving statistically significant improvement in alleviating forgetting compared to replaying random examples.

To summarize, the contributions of this paper are (1) empirical analysis on how forgotten examples are associated with learned tasks in representative 1B and 7B LLMs, and (2) a novel view of predicting example forgetting as a matrix completion problem, and (3) a practical algorithm to mitigate forgetting during LLM fine-tuning.

2 Problem and Experiment Setup

We define forgetting z_{ij} as degradation (increase) in log perplexity on an upstream example $x_j \in x_{1..N}$ after a LM learns a new task (set of examples) $T_i \in T_{1..M}$. We evaluate forgetting on N upstream examples when the model learns M tasks separately and record forgetting z_{ij} in a matrix $Z \in \mathbb{R}^{M \times N}$. We experiment with OLMo-1B, OLMo-7B and OLMo-7B-Instruct where pre-training data for language modeling and instruction-tuning is released.

OLMo-1B and 7B. OLMo models are pre-trained on Dolma [34], a massive collection of documents. We fine-tune LMs over separate tasks from FLAN-V2 and Tulu V2 instruction data [13], obtaining 77 fine-tuned models. We then evaluate log perplexity increase on a 1% subset of Dolma-v1.6-Sample. Each upstream example is a maximum 2,048-token document from Dolma, resulting in 141,816 examples. We used the same training configurations for fine-tuning 1B and 7B models.

OLMo-7B-Instruct. OLMo-7B-Instruct models are instruction-tuned on Tulu V2. In our experiments, we few-shot fine-tune OLMo-7B-Instruct over new task data from MMLU [9], BBH [35], and

Table 1: R^2 of fitting the association matrices Z of forgetting with simple models.

	OLMo-1B (Full)	OLMo-7B (Full)	OLMo-7B (LoRA)	OLMo-7B-Instruct (LoRA)
Upstream x only	0.5219	0.3070	0.3423	0.5837
Task T only	0.0835	0.1488	0.2700	0.0751
Additive Linear	0.5985	0.4518	0.6123	0.6588
Multiplicative	0.6083	0.4970	0.6537	0.7418

Table 2: Performance of predicting example forgetting. We report standard deviation over different sets of upstream examples with known ground truth forgetting ($\mathcal{S}$) beforehand.

Model	OLMo-1B	OLMo-7B	OLMo-7B (LoRA)	OLMo-7B-Instruct (LoRA)	
Task	FLAN	FLAN	FLAN	MMLU+BBH	
Metrics	RMSE (10^{-2}) ↓	RMSE (10^{-2}) ↓	RMSE (10^{-2}) ↓	RMSE (10^{-2}) ↓	F1 ↑
Additive	$2.81_{\pm 0.01}$	$7.40_{\pm 0.03}$	$3.50_{\pm 0.01}$	$6.12_{\pm 0.05}$	$54.83_{\pm 2.92}$
SVD	$2.88_{\pm 0.03}$	$7.53_{\pm 0.04}$	$3.49_{\pm 0.00}$	$6.24_{\pm 0.05}$	$51.91_{\pm 2.27}$
KNN	**2.79**$_{\pm 0.02}$	**7.33**$_{\pm 0.07}$	**3.45**$_{\pm 0.04}$	**5.54**$_{\pm 0.15}$	**70.52**$_{\pm 0.20}$
Rep-Dot	$3.84_{\pm 0.00}$	$9.29_{\pm 0.00}$	$5.45_{\pm 0.00}$	$6.19_{\pm 0.00}$	61.50 $_{\pm 0.00}$

TruthfulQA [21], and evaluate log perplexity increase over a stratified sample of 10,718 examples from Tulu v2 as upstream examples.

3 Associations between Learned and Forgotten Examples

We visualize the association matrices Z in Figure 1 for 4 different combinations of model and training setups (full parameter fine-tuning or LoRA). Each item z_{ij} denotes forgetting that happens on an upstream example x_j after fine-tuning on the new task T_i. The visualization indicates a mixture of simple (e.g. most columns being multiplication of the others) and more complicated patterns in associations. We quantitatively measure how well Z can be approximated with simple regression models with different inductive bias.

Models. We consider (1) additive linear models, where $z_{ij} = b + \alpha_i + \beta_j + \epsilon$, where α_i and β_i are learnable parameters associated with each new task or upstream example. (2) multiplicative models (SVD with rank r=1), where $z_{ij} = s\alpha_i\beta_j + \epsilon$.

Metrics. We measure R^2, a common metric for determining how well a regression model fits data. Let f_{ij} be the fitted value, R^2 is defined as $1 - \sum_{i,j}(z_{ij} - f_{ij})^2 / \sum_{i,j}(z_{ij} - \bar{Z})^2$.

Results. We report R^2 of different models in Table 1. In all the setups, the multiplicative model achieves better fit than additive models at the same number of trainable parameters. We note that multiplicative models are more suitable for situations where some upstream examples or new tasks almost never experience or inflict perplexity changes ($\alpha_i, \beta_j \approx 0$), which is indeed a predominant pattern from the visualization. The models achieve R^2 between 0.497 to 0.742 in different setups. Notably, OLMo-7B (Full) achieves a clearly lower R^2 of 0.497 than OLMo-1B (Full) of 0.608, indicating more complicated associations between learned tasks and forgotten examples for larger models. Similarly, on OLMo-7B, LoRA fine-tuning achieves a R^2 (of 0.654) higher than that of full fine-tuning, implying simpler associations compared to full fine-tuning.

In Appendix, we examine more complicated associations between learned tasks and forgotten examples with SVD of the association matrices Z.

4 Predicting Example Forgetting with Association Matrix Completion

Our analysis suggests a novel view of predicting example forgetting as a matrix completion problem. This is useful for targeted mitigation of forgetting as we replay predicted examples. Unlike prior works [14] that rely on an LM to encode contents of upstream examples and new tasks for prediction, we attempt to rely solely on example associations in Z without utilizing the contents in examples.

Our goal is to accurately predict forgetting z_{ij} over upstream examples $x_{1..M}$ when the model learns an unseen task without running expensive LLM inference on all $x_{1..M}$. To evaluate this, we create training and test splits by partitioning FLAN (OLMo) or MMLU and BBH (OLMo-Instruct) tasks. We assume knowing the ground truth forgetting z_{ij} of a tiny random set $\mathcal{S}$ ($|\mathcal{S}| = 30$) of upstream examples for a fine-tuned model, which takes seconds to obtain; our goal is to predict forgetting of the rest $10k - 100k$ upstream examples. Details such as train-test splits are discussed in Appendix C.

Table 3: Log perplexity over a held-out (never replayed) subset of upstream data from Dolma after full-parameter fine-tuning OLMo-7B on FLAN tasks. p-value is computed against replaying random examples (paired t-test on 20 tasks from the test split).

	Log PPL	p-value
Before Fine-Tuning	2.2787	-
No Replay	2.3511	-
Replay Random	2.3041	-
Replay w/ KNN	2.3016	0.024
Replay w/ Ground Truth	2.3007	0.005

Methods. We run matrix completion algorithms including additive linear, SVD, and k-nearest neighbors (KNN) models. We also compare with Rep-dot [14] which maps inner products of learned LM encoding of upstream and learned examples to forgetting.

Results of Predicting Example Forgetting. Table 2 summarizes the in-domain results of predicting example forgetting. We report Root Mean Squared Error (RMSE) of predicting the log-perplexity increase z_{ij}. For OLMo-Instruct, we additionally report F1 score of predicting whether $z_{ij} > 0$. We see KNN models consistently outperforms additive linear, SVD, and trainable representation-based prediction methods across different models and setups.

Mitigating Forgetting with Predicted Forgetting. We empirically verify the practical utility of predicting forgetting on OLMo-7B. As the model learns new tasks, we replay a small subset of upstream examples sparsely, prioritizing those with higher predicted forgetting, based on $\exp(z_{ij}/\tau)$ given by the KNN model, where τ is the temperature hyperparameter. Table 3 summarizes the log perplexity on a held-out subset of Dolma before and after full-parameter fine-tuning for measuring forgetting. As a reference, replaying with known ground truth forgetting achieves significantly ($p = 0.005$) lower log perplexity of 2.3007 compared to replaying random examples. Replaying with KNN predicted forgetting achieves log perplexity of 2.3016, significantly lower than random examples with $p = 0.024$.

5 Related Works

Related to our work, data attribution study faithful algorithms to find examples or tasks that account for a prediction [17, 12] when models are trained jointly on multiple examples or tasks. We instead focus on analysis of affected upstream example predictions when news tasks are learned. [5, 38, 4, 44, 36] identify memorized, important, or forgetful training data, but few analyze how these statistics depend on the learned tasks. [45, 43, 31] study prediction of task performance from training setups; we perform prediction at the example-level which is more fine-grained and under-explored. [26] study relationships between task similarity and forgetting in foundation models over a sequence of newly learned tasks; our work instead focus on forgetting of upstream data of LLMs. Prior works represented by [1, 40, 2] study selection strategies of examples for replay-based continual learning algorithms. We focus on analyzing patterns of forgetting and leave more comprehensive comparison to existing continual learning algorithms as future work.

6 Conclusions

In this paper, we empirically analyzed the associations between learned and forgotten examples in LM fine-tuning. We showed forgetting can be well-approximated with multiplicative effects of upstream and learned examples and visualized more complicated associations. We showed the example associations alone offer useful information to predict example forgetting when fine-tuning LMs on new tasks. We demonstrated practical utility of our analysis by showing reduced forgetting as we replay examples with predicted forgetting. We expect our results can inspire future study in a more practical online continual learning setup where tasks are learned sequentially.

References

[1] Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. *Neural Information Processing Systems*, 2019.

[2] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *Neural Information Processing Systems*, 2019.

[3] Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less. *ArXiv preprint*, abs/2405.09673, 2024.

[4] Stella Biderman, USVSN PRASHANTH, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[6] Daniel Goldfarb, Itay Evron, Nir Weinberger, Daniel Soudry, and PAul HAnd. The joint effect of task similarity and overparameterization on catastrophic forgetting — an analytical model. In *The Twelfth International Conference on Learning Representations*, 2024.

[7] Ian J. Goodfellow, Mehdi Mirza, Xia Da, Aaron C. Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[8] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, A. Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Daniel Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hanna Hajishirzi. Olmo: Accelerating the science of language models. *ArXiv preprint*, abs/2402.00838, 2024.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[11] Adam Ibrahim, Benjamin Th'erien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *ArXiv preprint*, abs/2403.08763, 2024.

[12] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. *ArXiv preprint*, abs/2202.00622, 2022.

[13] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew E. Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hanna Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *ArXiv preprint*, abs/2311.10702, 2023.

[14] Xisen Jin and Xiang Ren. What will my model forget? forecasting forgotten examples in language model refinement. In *International Conference on Machine Learning*, 2024.

[15] Damjan Kalajdzievski. Scaling laws for forgetting when fine-tuning large language models. *ArXiv preprint*, abs/2401.05605, 2024.

[16] Anat Kleiman, Jonathan Frankle, Sham M. Kakade, and Mansheej Paul. Predicting task forgetting in large language models. In *ICML 2023 Workshop DeployableGenerativeAI homepage*, 2023.

[17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, 2017.

[18] Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*, 2024.

[19] Sebastian Lee, Sebastian Goldt, and Andrew M. Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, 2021.

[20] Timothée Lesort, Oleksiy Ostapenko, Pau Rodríguez, Diganta Misra, Md Rifat Arefin, Laurent Charlin, and Irina Rish. Challenging common assumptions about catastrophic forgetting and knowledge accumulation. In *Conference on Lifelong Learning Agents*, pages 43–65. PMLR, 2023.

[21] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, 2022.

[22] Pratyush Maini, Saurabh Garg, Zachary Lipton, and J. Zico Kolter. Characterizing datapoints via second-split forgetting. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[23] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[24] Sanket Vaibhav Mehta, Darshan Patil, Sarath Chandar, and Emma Strubell. An empirical investigation of the role of pre-training in lifelong learning. *J. Mach. Learn. Res.*, 24:214:1–214:50, 2021.

[25] Seyed-Iman Mirzadeh, Arslan Chaudhry, Dong Yin, Huiyi Hu, Razvan Pascanu, Dilan Görür, and Mehrdad Farajtabar. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning, ICML*, 2022.

[26] Oleksiy Ostapenko, Timothée Lesort, Pau Rodr'iguez, Md Rifat Arefin, Arthur Douillard, Irina Rish, and Laurent Charlin. Continual learning with foundation models: An empirical study of latent replay. In *CoLLAs*, 2022.

[27] Haoran Que, Jiaheng Liu, Ge Zhang, Chenchen Zhang, Xingwei Qu, Yi Ma, Feiyu Duan, Zhiqi Bai, Jiakai Wang, Yuanxing Zhang, Xu Tan, Jie Fu, Wenbo Su, Jiamang Wang, Lin Qu, and Bo Zheng. D-cpt law: Domain-specific continual pre-training scaling law for large language models. *ArXiv*, abs/2406.01375, 2024.

[28] Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2021.

[29] Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[30] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In Vincent Y. Shen, Nobuo Saito, Michael R. Lyu, and Mary Ellen Zurko, editors, *Proceedings of the Tenth International World Wide Web Conference, WWW*, 2001.

[31] Viktoria Schram, Daniel Beck, and Trevor Cohn. Performance prediction via Bayesian matrix factorisation for multilingual natural language processing tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Dubrovnik, Croatia, 2023.

[32] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022.

[33] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ArXiv preprint*, abs/2404.16789, 2024.

[34] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Raghavi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, A. Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Daniel Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hanna Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. *ArXiv preprint*, abs/2402.00159, 2024.

[35] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *ArXiv preprint*, abs/2210.09261, 2022.

[36] Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, 2020.

[37] Mingxu Tao, Yansong Feng, and Dongyan Zhao. Can BERT refrain from forgetting on sequential tasks? a probing study. In *The Eleventh International Conference on Learning Representations*, 2023.

[38] Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. *ArXiv preprint*, abs/2205.10770, 2022.

[39] 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. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[40] Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, H. Lu, and Yujiu Yang. Inscl: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In *North American Chapter of the Association for Computational Linguistics*, 2024.

[41] Genta Winata, Lingjue Xie, Karthik Radhakrishnan, Shijie Wu, Xisen Jin, Pengxiang Cheng, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[42] Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. Continual learning for large language models: A survey. *ArXiv preprint*, abs/2402.01364, 2024.

[43] Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020.

[44] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *ArXiv preprint*, abs/2402.04333, 2024.

[45] Qinyuan Ye, Harvey Yiyun Fu, Xiang Ren, and Robin Jia. How predictable are large language model capabilities? a case study on big-bench. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

[46] Xiao Zhang and Ji Wu. Dissecting learning and forgetting in language model finetuning. In *The Twelfth International Conference on Learning Representations*, 2024.

[47] Haiyan Zhao, Tianyi Zhou, Guodong Long, Jing Jiang, and Chengqi Zhang. Does continual learning equally forget all parameters? In *International Conference on Machine Learning*, 2023.

A Discussions and More Related Works

In this paper, we primarily focused on analyzing the association between learned and forgotten examples. We list factors that are known to affect forgetting in prior study: (1) type and size of the LM [24, 32, 15, 25] (2) trainable parts of the model (*e.g.*, LoRA, soft prompts, or full-model tuning) [3, 29] (3) hyperparameters such as learning rate [11, 41], dropout [7], number of training steps [4, 16] (4) optimizer [20] and training algorithms (*e.g.*, various continual learning algorithms) [33, 42].

Mechanical Interpretation of Example Associations. We focused on empirical statistics of forgetting in this paper while treating the LLM as a black box. We believe research on mechanical interpretation of forgetting such as [37, 47, 18] is complementary to ours and can potentially explain in the future why the associations in Z are often simple, and in which circumstances the associations become more complicated.

Limitation of Replaying with Forgetting Prediction. We note that our current practice requires two runs of fine-tuning: a replay-free run of fine-tuning for which model the forgetting will be evaluated and predicted, and another run while replaying examples with the predicted forgetting. The practice is still efficient given the relative small training set of fine-tuning. In the continuing work, we will develop methods to predict forgetting on-the-fly during fine-tuning, mitigating the overhead of fine-tuning the model twice.

B Dataset, Model, and LM Training Details

Models. We use OLMo-7B[1] of the version pretrained on Dolma v1.6; and OLMo-7B-Instruct[2], which is tuned on Tulu v2 and other human feedback datasets.

Upstream Examples. We summarize the list of upstream datasets in Table 4. We also include the number of training examples in each dataset, initial log perplexity, and average forgetting occured on these datasets.

We examine forgetting on Dolma in our OLMo-1B and OLMo-7B experiments. We sample 1% of text chunks of length 2,048 from v1.6-sample version of the dataset, resulting in 141,816 chunks of length 2,048. We compute log perplexity over all 2,048 tokens in each example.

We examine forgetting on Tulu V2 in our OLMo-7B-Instruct experiments. We randomly sample an approximately balanced number of examples from each task in Tulu, and filter out examples with input length that exceeds 2,048 (the limit of OLMo models) after tokenization. This results in 10,718 examples. We compute log perplexity on ground truth output tokens only.

Learned New Tasks. We summarize the list of newly learned tasks in Tables 5 and 6 for OLMo-1B, 7B and OLMo-7B-Instruct experiments. We also include the number of training examples and forgetting caused by each task averaged over all upstream examples.

Training and Evaluation Details. For full-parameter fine-tuning of OLMo-1B and 7B, we train the model for 1,000 steps with an effective batch size of 8 and a linearly decaying learning rate of $2e^{-6}$. For LoRA fine-tuning, we set rank=64 in all our experiments and use a constant learning rate of 10^{-4}. We train the models for 625 steps with an effective batch size of 8. For OLMo-7B-Instruct and MMLU, BBH, TruthfulQA, considering the small size of the training sets, we train the models only for 37 steps with an effective batch size of 8. We use HuggingFace Transformers library for training and VLLM library for efficient inference. The statistics of forgetting are obtained in a single run.

Dataset Licenses and Safety. MMLU and BBH are released under MIT license. Truthful QA, Dolma, and OLMo models are released under Apache 2.0 license. Tulu V2 is released under ODC-By license. We thank [34] for removing personally identifiable information from the massive corpus, Dolma, before release as described in the original manuscript. The other datasets do not contain personally identifiable information according to our inspection.

[1] https://huggingface.co/allenai/OLMo-7B

[2] https://huggingface.co/allenai/OLMo-7B-Instruct

Table 4: Upstream tasks used in our experiments where forgetting is evaluated. We also report the number of training examples, log perplexity of these examples before fine-tuning (Init. Log PPL) and averaged forgetting happened on these examples averaged after learning different new tasks when performing LoRA fine-tuning on OLMo-7B models (Avg. Forgetting).

Task	#. Examples	Init. Log PPL	Avg. Forgetting
flan_v2	995	0.506	-0.006
cot	1000	0.347	-0.001
oasst1	1000	1.117	-0.043
lima	946	1.931	-0.045
gpt4_alpaca	1000	0.693	-0.018
code_alpaca	1000	0.402	-0.073
sharegpt	976	0.940	-0.049
wizardlm	1979	0.693	-0.025
open_orca	995	1.004	-0.080
science	687	0.322	-0.021
hard_coded	140	2.682	-0.353
Dolma	141816	2.283	0.035

C Details of Forgetting Prediction and Replay

Data Splits for Predicting Example Forgetting. We mark the tasks used as in-domain test splits for predicting example forgetting (Sec. 4) in Tables 5 and 6.

Training and Evaluation Details. We use Surprise Library 1.1.3[3] for additive linear, SVD, and KNN prediction models. For SVD, we set the dimension of the learnable features as 5. KNN aggregates the forgetting of other upstream examples based on the similarity between forgetting patterns of a seen task and the unseen task over a small subset with known ground truth forgetting.

For in-domain test splits, we randomly sample 30 upstream examples and assume the ground truth forgetting is known for these examples. This is required for predicting forgetting on the rest of upstream examples by additive linear, SVD, and KNN methods. We repeat the experiment 10 times and report the mean and standard deviation in Table 2.

We used OLMo-1B models as the trainable example encoders in the implementation of the prediction method by [14] (Rep-dot) that relies on inner products of trained example representations. We notice these models trained with mean squared error objective perform poorly on F1 metrics. Therefore, for F1 metrics reported for Rep-dot in Table 2, we used a variant using cross-entropy as the optimization objective. At inference, given an upstream example, we compute the averaged dot-product with all examples in the learned task. We note that at inference time Rep-dot does not require ground truth forgetting of a small number of examples. For a fair comparison with other matrix completion methods, we replace the prediction of Rep-dot with ground truth forgetting on these examples.

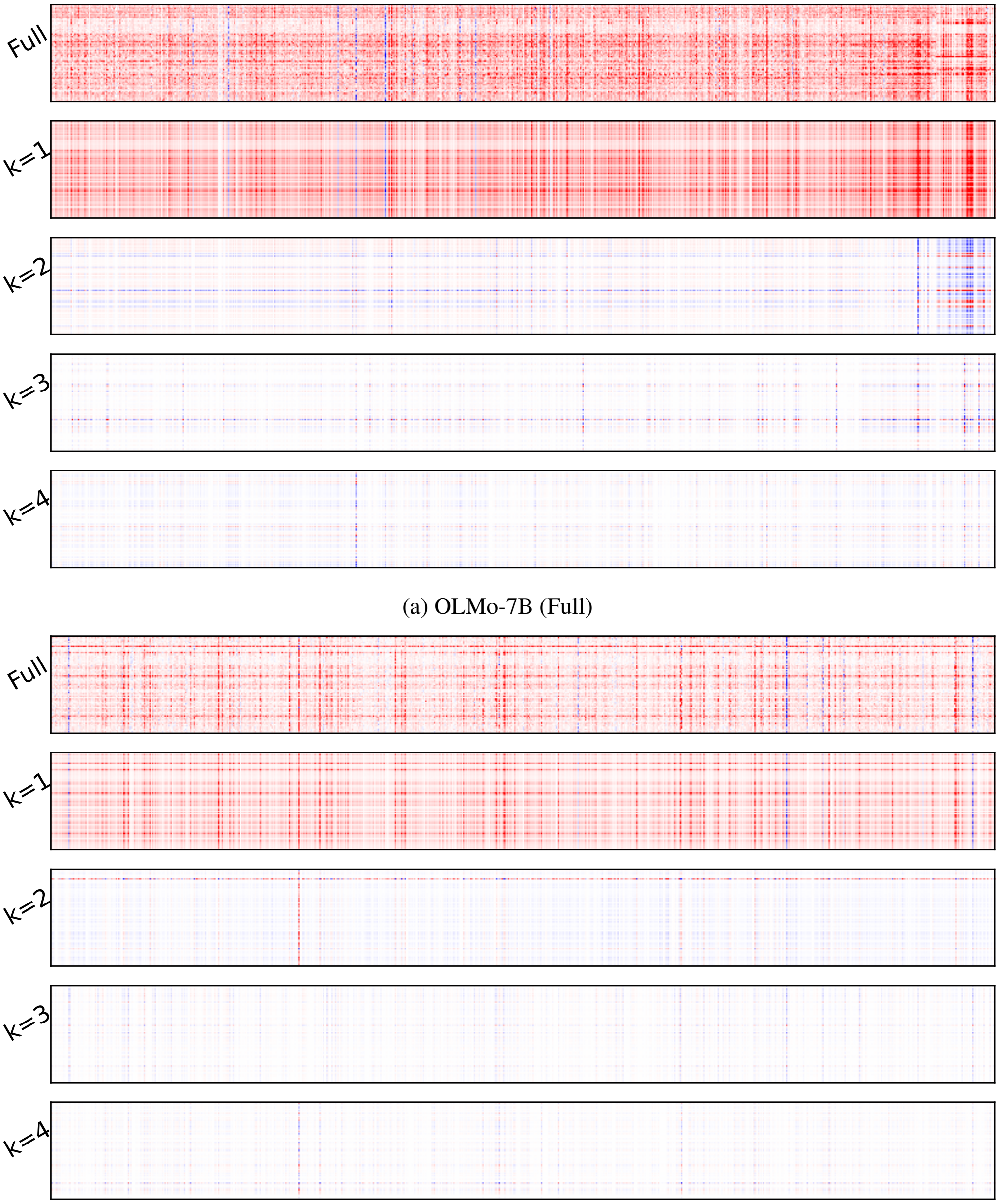
Replaying Upstream Examples in Fine-Tuning. We sparsely replay 1 mini-batch of 8 upstream examples every 32 steps of model update while fine-tuning on new tasks. Given predicted or ground truth forgetting $z_{i,1..J}$ on upstream examples $x_{1..J}$ when learning a new task T_i, we sample upstream examples to replay from a categorical distribution where $p(x_j) \propto \exp(z_{i,j}/\tau)$, where τ is a temperature hyperparameter set as 0.1. The hyperparameter τ is tuned on a single validation task by using ground truth forgetting Z.

D Additional Results about Example Associations

We visualize progressive reconstruction with k-th singular value and singular vectors for OLMo experiments in Figure 2. We see when $k = 2$, there is a single row and column with significantly larger forgetting than the others. This pattern exemplifies a complicated association that is not captured by the simple multiplicative model ($k = 1$).

We further show the distribution of singular values and R^2 of reconstruction of Z in our OLMo and OLMo-Instruct experiments in Figure 4.

[3] https://github.com/NicolasHug/Surprise/tree/v1.1.3



(a) OLMo-7B (Full)

(b) OLMo-7B (LoRA)

Figure 2: Reconstruction of Z in OLMo-7B experiments with k-th singular value and vectors. Higher values of k capture finer-grained details in Z.

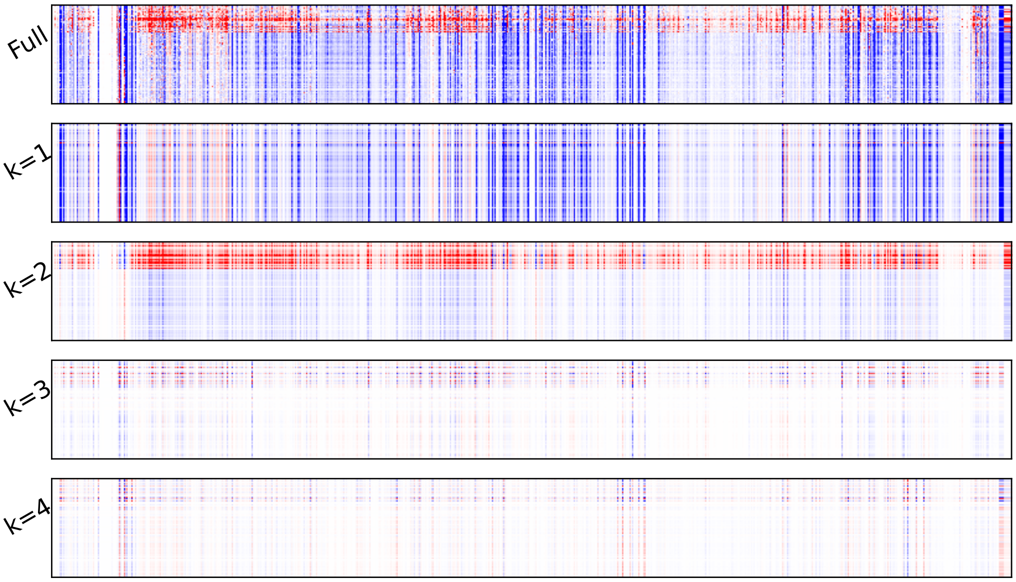
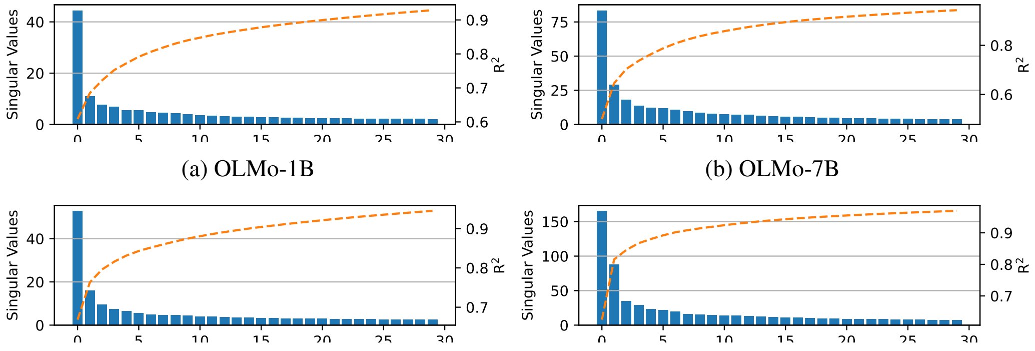


Figure 3: Reconstruction of Z in OLMo-7B-Instruct experiments with k-th singular value and vectors. Higher values of k capture finer-grained details in Z.



(a) OLMo-1B

(b) OLMo-7B

(c) OLMo-7B (LoRA)

(d) OLMo-7B-Instruct (LoRA)

Figure 4: Singular values (bars) and R^2 (dash lines) of reconstruction of Z with up to k-th singular value and singular vectors.

Task Category	Task	#. Examples	Avg. Forgetting Caused	Task Category	Task	#. Examples	Avg. Forgetting Caused
FLAN V2	aeslc	28860	0.012		quac	60000	0.070
	ag_news_subset	60000	0.020		record	60000	0.030
	anli_r1	33880	0.018		rte	4580	0.034
	anli_r2	60000	0.022		samsum	29460	0.016
	anli_r3	60000	0.023		sentiment140	60000	0.012
	arc_challenge*	1820	0.027		snli	59900	0.027
	arc_easy*	4080	0.027		squad_v1	60000	0.067
	bool_q	18440	0.056		squad_v2	60000	0.148
	cb*	400	0.037		sst2	60000	0.024
	cnn_dailymail	60000	0.007		story_cloze	3340	0.032
	cola	16700	0.040		stsb	11280	0.031
	common_gen	60000	0.037		trec*	10500	0.026
	copa	700	0.043		trivia_qa*	60000	0.018
	coqa*	14180	0.134		true_case	58520	0.053
	cosmos_qa	50120	0.046		web_nlg_en	60000	0.048
	dart	60000	0.033		wic*	10440	0.046
	definite_pronoun_resolution*	2240	0.015		wiki_lingua_english_en	60000	0.015
	drop	60000	0.045		wmt14_enfr	60000	0.017
	e2e_nlg*	60000	0.046		wmt16_translate_csen	60000	0.009
	fix_punct*	56140	0.046		wmt16_translate_deen	60000	0.011
	gigaword	32240	0.011		wmt16_translate_fien	60000	0.013
	glue_mrpc	6920	0.059		wmt16_translate_roen	60000	0.015
	glue_qqp*	60000	0.032		wmt16_translate_ruen*	60000	0.014
	hellaswag	60000	0.027		wmt16_translate_tren*	60000	0.017
	imdb_reviews	49600	0.013		wnli	1200	0.024
	math_dataset*	60000	0.043		word_segment	60000	0.107
	mnli_matched	60000	0.057		wsc*	1000	0.016
	mnli_mismatched	60000	0.066		yelp_polarity_reviews*	60000	0.013
	multi_news	60000	0.010	Tulu	open_orca	29683	0.009
	multirc	54080	0.058		oasst1	7331	0.005
	natural_questions*	60000	0.010		lima	1018	0.194
	openbookqa*	9900	0.046		code_alpaca	20016	0.015
	opinion_abstracts_idebate*	3300	0.024		gpt4_alpaca	19906	0.016
	opinion_abstracts_rotten_tomatoes	6260	0.008		cot	49747	0.019
	para_crawl_enes	60000	0.018		science	7468	0.022
	paws_wiki	60000	0.063		sharegpt	111912	0.010
	piqa	32020	0.037		hard_coded	140	0.056
	qnli*	60000	0.043		wizardlm	29810	0.019

Table 5: The list of learned tasks in our experiments on OLMo-7B. We also include the number of training examples in each task (#. Example) and forgetting caused by each learned task after LoRA fine-tuning averaged over all upstream examples as a reference. * notes for tasks used as the in-domain test split in forgetting prediction experiments in Sec. 4.

Task Category	Task	#. Examples	Avg. Forgetting Caused	Task Category	Task	#. Examples	Avg. Forgetting Caused
MMLU	abstract_algebra	11	-0.030	BBH	boolean_expressions*	125	-0.095
	anatomy	14	-0.076		causal_judgement	93	-0.119
	astronomy	16	-0.074		date_understanding	125	-0.122
	business_ethics	11	-0.042		disambiguation_qa	125	-0.086
	clinical_knowledge	29	-0.093		dyck_languages*	125	-0.090
	college_biology*	16	-0.069		formal_fallacies*	125	-0.087
	college_chemistry	8	-0.088		geometric_shapes	125	-0.045
	college_computer_science	11	-0.057		hyperbaton*	125	-0.093
	college_mathematics	11	-0.065		logical_deduction_five_objects*	125	-0.092
	college_medicine*	22	-0.072		logical_deduction_seven_objects	125	-0.089
	college_physics	11	-0.058		logical_deduction_three_objects	125	-0.116
	computer_security	11	-0.080		movie_recommendation*	125	-0.068
	conceptual_physics*	26	-0.087		multistep_arithmetic_two	125	-0.081
	econometrics	12	-0.043		navigate	125	-0.067
	electrical_engineering	16	-0.108		object_counting*	125	-0.106
	elementary_mathematics	41	-0.098		penguins_in_a_table	73	-0.122
	formal_logic	14	-0.039		reasoning_about_colored_objects	125	-0.134
	global_facts*	10	-0.012		ruin_names	125	-0.098
	high_school_biology*	32	-0.098		salient_translation_error_detection	125	-0.077
	high_school_chemistry	22	-0.096		snarks	89	-0.125
	high_school_computer_science	9	-0.059		sports_understanding	125	-0.113
	high_school_european_history*	18	-0.084		temporal_sequences	125	-0.099
	high_school_geography	22	-0.090		tracking_shuffled_objects_five_objects	125	-0.049
	high_school_government_and_politics	21	-0.057		tracking_shuffled_objects_seven_objects	125	-0.100
	high_school_macroeconomics	43	-0.111		tracking_shuffled_objects_three_objects	125	-0.091
	high_school_mathematics	29	-0.075		web_of_lies	125	-0.066
	high_school_microeconomics	26	-0.070		word_sorting	125	-0.124
	high_school_physics*	17	-0.067	TruthfulQA	Nutrition	16	0.078
	high_school_psychology	60	-0.085		Stereotypes	24	-0.044
	high_school_statistics	23	-0.063		Confusion	46	-0.109
	high_school_us_history*	22	-0.085		Psychology	19	0.018
	high_school_world_history	26	-0.070		Language	21	-0.058
	human_aging*	23	-0.046		Sociology	55	-0.147
	human_sexuality*	12	-0.081		Finance	9	0.096
	international_law	13	-0.050		Indexical Error	57	-0.107
	jurisprudence	11	-0.088		Science	9	0.171
	logical_fallacies*	18	-0.024		Misconceptions	104	-0.124
	machine_learning	11	-0.072		Economics	31	-0.061
	management*	11	-0.062		Education	10	0.103
	marketing*	25	-0.043		Proverbs	18	-0.029
	medical_genetics	11	-0.029		Conspiracies	25	0.058
	miscellaneous	86	-0.128		Religion	15	-0.029
	moral_disputes	38	-0.101		Statistics	5	0.240
	moral_scenarios*	100	-0.068		Misquotations	16	0.077
	nutrition	33	-0.098		Subjective	9	-0.017
	philosophy*	34	-0.063		Law	64	-0.125
	prehistory	35	-0.093		History	24	-0.026
	professional_accounting	31	-0.075		Fiction	30	-0.096
	professional_law	170	-0.157		Mandela Effect	6	0.008
	professional_medicine*	31	-0.087		Politics	10	-0.037
	professional_psychology	69	-0.117		Misinformation	12	-0.030
	public_relations*	12	-0.073		Logical Falsehood	14	-0.028
	security_studies	27	-0.109		Distraction	14	-0.091
	sociology*	22	-0.075		Weather	17	0.006
	us_foreign_policy*	11	-0.050		Myths and Fairytales	21	0.068
	virology	18	-0.079		Superstitions	22	-0.064
	world_religions	19	-0.049		Advertising	13	-0.078
					Paranormal	26	-0.074
					Health	55	-0.137

Table 6: The list of learned tasks in our experiments on OLMo-7B-Instruct. We include the number of training examples in each task (#. Examples), and forgetting caused by each learned task after LoRA fine-tuning averaged over all upstream examples (Avg. Forgetting Caused) as a reference. * notes for tasks used as the in-domain test split in forgetting prediction experiments in Sec. 4.