Demystifying Language Model Forgetting with Low-rank Example Associations

Xisen Jin [1] **Xiang Ren** [1]

Abstract

Unintended forgetting of learned knowledge imposes an inherent risk on fine-tuning LLMs with new task data. In this paper, we adapt a classical framework for risk management to address forgetting by predicting their occurrence and perform targeted mitigation. Specifically, we show empirically that the associations between learned tasks and forgotten upstream examples are often well-approximated with low-rank matrices in diverse setups. Leveraging such empirical associations collected in the past, we predict forgetting of upstream examples when fine-tuning on unseen tasks with matrix completion. At the fine-tuning stage, we upweight examples predicted to be more forgotten for replay and demonstrate statistically significantly reduced forgetting over held-out, never-replayed upstream data. In summary, we show that predicting forgetting with low-rank example associations enables efficient examination and mitigation of forgetting.

1. Introduction

There has been a growing need for continued fine-tuning of LLMs to mitigate harmful behaviors, update outdated knowledge, and adapt to unseen tasks and domains. Although fine-tuning allows efficient and incremental adaptation of models, models may suffer from catastrophic forgetting (McCloskey & Cohen, 1989; Goodfellow et al., 2014) of upstream knowledge learned in the pre-training or instruction-tuning phase, causing unintended prediction changes over known information. This is problematic for the performance and reliability of online deployed LLMs, limiting the feasibility of continual fine-tuning in practice (Raffel, 2023; Shi et al., 2024).

In this paper, we borrow a classical risk management framework in systems and software engineering (Boehm, 1991) to address forgetting; we predict the likelihood of risks (*i.e.,* example forgetting) from past experiences (*i.e.,* past fine-tuning runs) and apply targeted mitigation. We assume the same LLM is independently fine-tuned on diverse custom datasets, for example, by an LLM fine-tuning service provider. We illustrate the setup in Figure 1. We explore how forgetting caused by unseen tasks can be efficiently predicted, and more specifically, from the forgetting that occurred while learning other tasks.

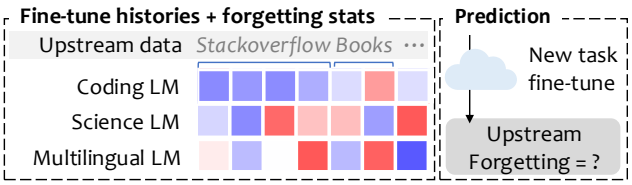


Figure 1. We consider the problem of fine-tuning LLMs on diverse custom datasets while reducing unintended forgetting, which is important, for example, to LLM service providers that offer fine-tuning APIs. Our research explores prediction of upstream data forgetting when fine-tuning on new tasks by leveraging forgetting statistics collected in the past. Compared to evaluating forgetting over the entire upstream data for each new fine-tuning task, predicting forgetting can enable more efficient examination and mitigation of forgetting.

Specifically, we start by examining associations between the learned tasks and forgotten upstream examples in LLM fine-tuning. This is because the complexity of such associations plays an important role in predictability; Figure 2 (a) illustrates a hypothetical scenario where certain upstream examples suffer more forgetting regardless of the learned tasks, making forgetting easily predictable; in contrast, (b) exemplifies upstream example forgetting that is highly dependent on the learned tasks. We measure forgetting (in continuous log perplexity increase or binary exact match drop) over N upstream examples, after fine-tuning the model on one of M unseen instruction-tuning tasks, while summarizing the results in a $M \times N$ matrix. We evaluate the complexity of the associations by measuring the goodness-of-fit of low-rank approximations of the matrix.

Our findings suggest that the associations between learned tasks and forgotten examples are often well-approximated

[1]University of Southern California. Correspondence to: Xisen Jin <xisenjin@usc.edu>.

ICML 2025 Workshop on Efficient Systems for Foundation Models (ES-FoMo), Vancouver, Canada. Copyright 2025 by the author(s).

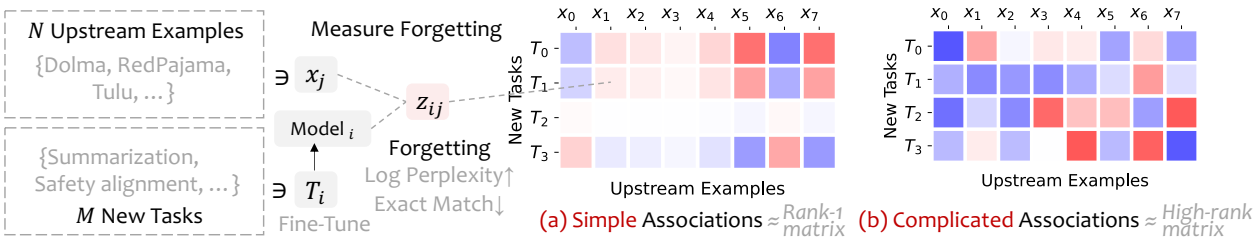


Figure 2. Analyzing the associations between learned tasks and forgotten upstream examples as we fine-tune LLMs on one of unseen new tasks. Over total N upstream examples and M unseen tasks, we measure and record forgetting (in red) in a $M \times N$ matrix and attempt to fit the associations with low-rank approximations. Better approximations of low-rank approximations indicate simpler associations between learned tasks and forgotten upstream examples.

with low-rank matrices across various model types (OLMo, OLMo2, MPT, Pythia) and sizes (1B to 13B parameters). For example, on OLMo-1B and OLMo-7B, rank-3 approximation fits the associations between 85 learned tasks (from FLAN, Tulu, and Dolly) and 140,000 upstream examples (from Dolma) with $R^2 > 0.69$.

Following the low-rank approximations of the associations, we predict example forgetting on unseen tasks by solving a matrix completion problem over the association matrices, analogous to collaborative filtering (Sarwar et al., 2001) in recommender systems, achieving both efficiency and interpretability. As an example, we achieve 58.16 F1 in predicting binary example forgetting with matrix factorization (MF) where the F1 of random guess is only 6.4.

Lastly, we mitigate forgetting with the predicted forgetting. We upweight upstream examples with higher predicted forgetting during replay as we fine-tune LLMs on new instruction-tuning tasks, achieving statistically significantly reduced over held-out, never-replayed upstream examples compared to replaying random examples.

To summarize, the contributions of this paper are (1) an empirical analysis on how forgotten examples are associated with learned tasks in representative 1B to 13B language models, and (2) a novel approach of predicting example forgetting by solving a matrix completion problem over the empirical associations, and (3) a scalable and efficient algorithm to mitigate forgetting during LLM fine-tuning by upweighting upstream examples for replay according to the predicted forgetting.

2. Problem and Experiment Setup

We start by defining the forgetting metrics and introduce the models and datasets studied in the experiments. We then briefly summarize our findings about the associations between learned tasks and forgotten upstream examples.

2.1. Collecting Statistics of Forgetting

We fine-tune an LLM (or an instruction-tuned LLM) on one unseen instruction-tuning task T_i from a collection of tasks $T_{1..M}$. This results in M separately fine-tuned models $f_{1..M}$. We then evaluate performance degradation on each upstream example $x_j \in x_{1..N}$ as forgetting. We use log perplexity as the main performance metric as it is applicable to both language modeling and instruction tuning, and is known to correlate well with other dataset-specific metrics (Hoffmann et al., 2022). For instruction tuning tasks with a restricted output space (*e.g.,* multi-choice questions), we also measure binary exact matches (EM). We measure forgetting z_{ij} that occurs on an upstream example $x_j \in x_{1..N}$ as increase in log perplexity or drop of exact match after fine-tuning the LM on a new task $T_i \in T_{1..M}$. We record forgetting z_{ij} in an association matrix $Z \in \mathbb{R}^{M \times N}$.

2.2. Models and Datasets

Our experiments require open access to upstream data of LLMs. We perform main experiments with OLMo-1B, OLMo-7B, and OLMo-7B-Instruct (Groeneveld et al., 2024). We also include OLMo2-7B, OLMo2-13B (OLMo et al., 2024), MPT-7B (Computer, 2023), and Pythia-1B to 12B (Biderman et al., 2023) for studying the complexity of the example associations across model types and sizes.

Upstream examples $x_{1..N}$ where forgetting is evaluated. For OLMo, OLMo2, MPT, and the Pythia family, we evaluate log perplexity increase over Dolma (Soldaini et al., 2024), OLMo2-Mix (OLMo et al., 2024), Redpajama (Computer, 2023), and Pile (Gao et al., 2020) respectively, each corresponding to their upstream pretraining corpora. We sample $10k$ to $140k$ documents truncated into 2,048 tokens. For OLMo-Instruct, we evaluate log perplexity increase on Tulu V2 (Ivison et al., 2023) which the model is instruction-tuned on. For the FLAN (Longpre et al., 2023) subset of Tulu, we also measure the drop of binary exact matches over correctly predicted upstream examples before fine-tuning.

Unseen Task $T_{1..M}$ where models are fine-tuned. We

fine-tune non-instruction-tuned models over 66 tasks from FLAN, 11 tasks from Tulu and 8 tasks from Dolly (Conover et al., 2023). For OLMo-Instruct, we fine-tune OLMo-7B-Instruct over new task data from MMLU (Hendrycks et al., 2021), BBH (Suzgun et al., 2022), TruthfulQA (Lin et al., 2022), Dolly, and OLMo2-SFT-Mix (OLMo et al., 2024).

We fine-tune full model parameters with a 2e-6 learning rate and other consistent configurations. Training details are included in Appendix D.

2.3. Empirical low-rank approximations of forgetting

Due to space limits, we leave detailed analysis in Appendix B. The key findings are that (1) low-rank example associations are prevalent in LLM fine-tuning in the models and training configurations examined. The association becomes simpler for bad learning rates, larger batch sizes, LoRA fine-tuning, and becomes (relatively) more complicated for larger and more capable models (e.g. OLMo2-7B compared to OLMo-7B). We consider this conclusion to hold for a collection of fine-tuning tasks with a similar level of heterogeneity. (2) Content-based similarities of learned tasks and upstream examples, such as token or embedding similarity, correlate poorly with example forgetting. These findings motivate us to predict upstream example forgetting with past forgetting statistics.

3. Predicting and Mitigating Example Forgetting

3.1. Training and Evaluation of Forgetting Prediction

Our first goal is to accurately predict forgetting z_{ij} over upstream examples $x_{1..M}$ when the model is fine-tuned on an unseen task T_j with a prediction model g, without running expensive LLM inference on all $x_{1..M}$. To evaluate this, we create training and test splits by partitioning the set of fine-tuning tasks (noted as $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{test}}$) and the rows of the association matrices Z. We further control whether $\mathcal{T}_{\text{train}}$ and $\mathcal{T}_{\text{test}}$ belong to the same category of tasks to test both in-domain and out-of-domain generalization ability of the prediction models. For OLMo-1B and 7B experiments, we use FLAN as in-domain tasks and Tulu and Dolly as out-of-domain testing tasks. For OLMo-7B-Instruct experiments, we use MMLU, BBH, OLMo2-SFT-Mix as in-domain tasks and use TruthfulQA and Dolly as out-of-domain testing tasks. Details about the tasks included in the training, in-domain testing, and out-of-domain testing sets are discussed in Tables 10 and 11 in Appendix E.

Matrix completion based prediction. We run matrix completion algorithms to predict forgetting after fine-tuning a model on a new task from the association matrix Z. We examine additive linear models (Additive), matrix factorization (MF), and k-nearest neighbors (KNN) models.

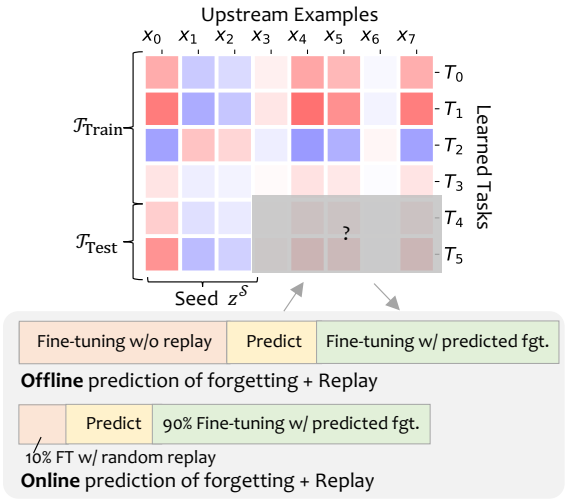


Figure 3. The training and testing setup of predicting example forgetting with association matrix completion, and their integration into example replay methods to mitigate forgetting.

Content based prediction baseline. We encode upstream examples x_j and the learned examples $x_i^{1..N_i} \in T_i$ with a pretrained frozen transformer sentence embedding model $h(\cdot)$ followed by two trainable MLP layers to obtain their representations (Reimers & Gurevych, 2019). The final prediction is made with the inner products of two representations $\langle h(x_j), \frac{1}{N_i}\sum_{N_i} h(x_i)\rangle$.

We leave the implementation details of matrix completion approaches and the sentence embedding approach in Appendix E.

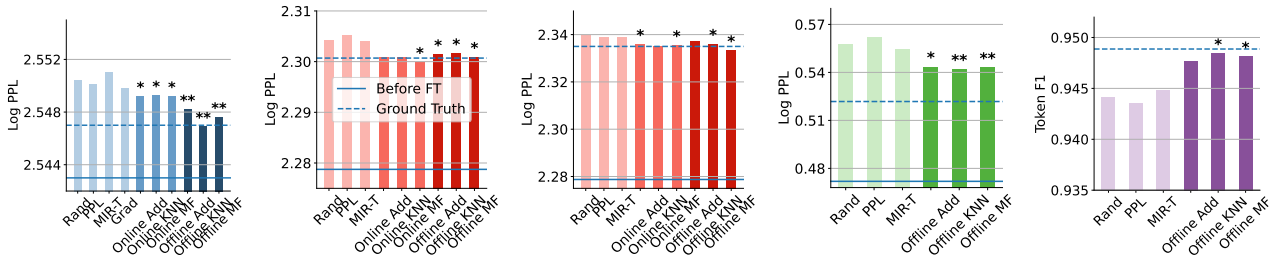
3.2. Mitigating Forgetting with Predicted Forgetting

Leveraging predicted forgetting for mitigating forgetting. We examine the practical utility of predicting forgetting as we sparsely replay upstream examples during forgetting following Jin & Ren (2024); de Masson D'Autume et al. (2019); Ibrahim et al. (2024). We perform targeted mitigation of forgetting by prioritizing examples that are predicted to suffer more from forgetting. This is achieved with weighted sampling of upstream examples x_j proportional to $\exp(\hat{z}_{ij}/\tau)$, where $\hat{z}_{ij}$ is the predicted forgetting and τ is a temperature hyperparameter.

We consider an offline and an online variant of replaying with predicted forgetting. We illustrate the two variants in Figure 3. The *offline* variant performs a replay-free run of fine-tuning on the task T_i, after which the forgetting over a small "seed" set of upstream examples will be evaluated. The forgetting over the rest upstream examples are predicted with matrix completion algorithms from the seed forgetting. We then perform another run of fine-tuning while replaying

Table 1. RMSE (↓) or F1(↑) of predicting example forgetting over a held-out set of upstream examples after fine-tuning LMs on unseen new tasks. We report average performance over 10 random seed sets ($\mathcal{S}$) of upstream examples with known ground truth forgetting beforehand.

	In-Domain					Out-of-Domain				
Model	OLMo-1B	OLMo-7B	MPT	OLMo-7B Instruct		OLMo-1B	OLMo-7B	MPT	OLMo-7B Instruct	
Metrics	RMSE	RMSE	RMSE	RMSE	F1	RMSE	RMSE	RMSE	RMSE	F1
Additive	2.81	7.40	13.33	15.57	55.81	**2.81**	5.83	10.02	38.90	43.57
KNN	**2.79**	7.33	12.80	14.30	56.87	2.84	5.83	7.71	**38.77**	**44.11**
MF	2.80	**7.14**	**10.41**	**13.74**	**58.16**	2.82	**5.76**	**7.03**	40.47	42.91
Embedding	2.81	7.44	13.86	13.94	55.46	3.53	6.16	10.59	41.22	42.95



(a) 1B, FLAN PPL (b) 7B, FLAN PPL (c) 7B, Tulu&Dolly PPL (d) $7B_{Ins}$, Dolly PPL (e) $7B_{Ins}$, OLMo2Mix F1

Figure 4. Log perplexity (↓) or Token F1 (↑) over upstream data by replay example selection strategies. The solid horizontal lines indicate the log perplexity before fine-tuning (*i.e.,* no forgetting). The dash lines show the results achieved by upweighting upstream examples according to their ground truth forgetting. * and ** indicate statistical significance of improvement ($p < 0.05$ or $p < 0.005$) compared to replaying random examples in paired t-tests on all fine-tuning tasks.

examples with the predicted forgetting. This creates computational overhead equivalent to one extra run of fine-tuning, but is still efficient when the training set of fine-tuning is considerably smaller than the upstream data. The *online* variant instead replays random examples for the first 10% of the fine-tuning steps, after which it evaluates seed forgetting and determine examples to be replayed in the rest of the 90% steps. Compared to the offline variant, this mitigates the extra overhead of fine-tuning by trading off the prediction accuracy of forgetting.

Evaluation and Metrics. We measure log-perplexity increase or Token F1 (a softer metric than exact match (Rajpurkar et al., 2016)) drop over a held-out subset of 10,000 examples from the upstream data. This ensures none of the test examples are selected for replay by any of the example selection strategies. We primarily examine whether weighted sampling with predicted forgetting statistically significantly improves over random sampling of upstream examples (Rand). We compare with alternative selection strategies of upstream examples for sparse replay (PPL, MIR-T, GradProd) and detail the approaches in Appendix E.

3.3. Results

Results of predicting example forgetting. Table 1 summarizes the error of predicting example forgetting over the tasks from the in-domain and out-of-domain test splits. We see matrix completion approaches consistently outperform the sentence embedding model adapted from the prior work (Jin & Ren, 2024). Among the three matrix completion approaches, we notice that MF models in general achieve the lowest prediction error. Besides, KNN in general outperforms additive linear and the embedding model while being highly computationally efficient.

Mitigating forgetting with the predicted forgetting. We leverage the online or offline predicted forgetting by the matrix completion algorithms to reweigh examples during replay following the procedure introduced in Sec. 3.2. Figure 4 summarizes log perplexity or Token F1 over the held-out (never replayed) upstream data as we apply different upstream example selection approaches. Upweighting examples with ground truth forgetting (GT) consistently reduces forgetting compared to random examples, shown in dash lines in Figure 4. By utilizing predicted forgetting by offline additive linear, KNN, and MF models, we statistically significantly reduce forgetting compared to random examples. The MF model achieves statistical significance in most of the scenarios, which aligns with its top average prediction performance of example forgetting in Table 1. Utilizing online-predicted forgetting also statistically significantly improves over replaying random examples in 3 of

the setups.

Further discussions. We show in Appendix G that targeted replay also slightly improves new task learning perplexity. We evaluate fine-tuned OLMo-7B-Instruct on unseen LLM leaderboard tasks and present the results in Appendix G. Although we observe slightly improved performance of replaying predicted examples to random or no replay on most forgotten tasks, we do not see statistical significance. Future works can mitigate downstream task forgetting with predicted forgetting with alternative algorithms that leverage past examples (Aljundi et al., 2017; Buzzega et al., 2020). Appendix I demonstrates mild change in the pattern of forgetting in the held-out upstream examples when replay is applied. Our results in Appendix J show that matrix completion and embedding based methods can be combined to better improve forgetting prediction performance.

4. Conclusions

In this paper, we empirically analyzed the associations between learned and forgotten examples in LM fine-tuning. We showed the example associations alone offer useful information to predict example forgetting when fine-tuning LMs on new tasks. We demonstrated the practical utility of our analysis by showing reduced forgetting as we reweigh examples for replay with predicted forgetting. We discuss limitations of the work in Appendix A.

References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., and Tuytelaars, T. Memory aware synapses: Learning what (not) to forget. *ArXiv*, abs/1711.09601, 2017.

Aljundi, R., Caccia, L., Belilovsky, E., Caccia, M., Lin, M., Charlin, L., and Tuytelaars, T. Online continual learning with maximally interfered retrieval. *Neural Information Processing Systems*, 2019a.

Aljundi, R., Lin, M., Goujaud, B., and Bengio, Y. Gradient based sample selection for online continual learning. In *Neural Information Processing Systems*, 2019b.

Biderman, D., Ortiz, J. G., Portes, J., Paul, M., Greengard, P., Jennings, C., King, D., Havens, S., Chiley, V., Frankle, J., Blakeney, C., and Cunningham, J. P. Lora learns less and forgets less. *ArXiv preprint*, abs/2405.09673, 2024a.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Biderman, S., PRASHANTH, U., Sutawika, L., Schoelkopf, H., Anthony, Q., Purohit, S., and Raff, E. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36, 2024b.

Boehm, B. W. Software risk management: principles and practices. *IEEE software*, 8(1):32–41, 1991.

Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. Dark experience for general continual learning: a strong, simple baseline. *ArXiv*, abs/2004.07211, 2020.

Chaudhry, A., Ranzato, M., Rohrbach, M., and Elhoseiny, M. Efficient lifelong learning with a-GEM. In *International Conference on Learning Representations*, 2019.

Computer, T. Redpajama: an open dataset for training large language models, 2023. URL https://github.com/togethercomputer/RedPajama-Data.

Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-

de Masson D'Autume, C., Ruder, S., Kong, L., and Yogatama, D. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Ding, M., Ji, K., Wang, D., and Xu, J. Understanding forgetting in continual learning with linear regression. In *Forty-first International Conference on Machine Learning*, 2024.

Doan, T. V., Bennani, M. A., Mazoure, B., Rabusseau, G., and Alquier, P. A theoretical analysis of catastrophic forgetting through the ntk overlap matrix. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Evron, I., Moroshko, E., Ward, R., Srebro, N., and Soudry, D. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pp. 4028–4079. PMLR, 2022.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Ganguli, D., Hernandez, D., Lovitt, L., Dassarma, N., Henighan, T., Jones, A., Joseph, N., Kernion, J., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Elhage, N., Showk, S. E., Fort, S., Hatfield-Dodds, Z., Johnston, S., Kravec, S., Nanda, N., Ndousse, K., Olsson, C., Amodei, D., Amodei, D., Brown, T. B., Kaplan, J., McCandlish, S., Olah, C., and Clark, J. Predictability and surprise in large generative models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Goodfellow, I. J., Mirza, M., Da, X., Courville, A. C., and Bengio, Y. An empirical investigation of catastrophic forgeting in gradient-based neural networks. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Groeneveld, D., Beltagy, I., Walsh, P., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K. R., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., Khot, T., Merrill, W., Morrison, J. D., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Pyatkin, V., Ravichander, A., Schwenk, D., Shah, S., Smith, W., Strubell, E., Subramani, N., Wortsman, M., Dasigi, P., Lambert, N., Richardson, K., Zettlemoyer, L., Dodge, J., Lo, K., Soldaini, L., Smith, N. A., and Hajishirzi, H. Olmo: Accelerating the science of language models. *ArXiv preprint*, abs/2402.00838, 2024.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models. *ArXiv*, abs/2203.15556, 2022.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. 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.

Huang, J., Cui, L., Wang, A., Yang, C., Liao, X., Song, L., Yao, J., and Su, J. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Annual Meeting of the Association for Computational Linguistics*, 2024.

Ibrahim, A., Th'erien, B., Gupta, K., Richter, M. L., Anthony, Q., Lesort, T., Belilovsky, E., and Rish, I. Simple and scalable strategies to continually pre-train large language models. *ArXiv preprint*, abs/2403.08763, 2024.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., and Madry, A. Datamodels: Predicting predictions from training data. *ArXiv preprint*, abs/2202.00622, 2022.

Ivison, H., Wang, Y., Pyatkin, V., Lambert, N., Peters, M. E., Dasigi, P., Jang, J., Wadden, D., Smith, N. A., Beltagy, I., and Hajishirzi, H. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *ArXiv preprint*, abs/2311.10702, 2023.

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

Johnson, W. B. and Lindenstrauss, J. Extensions of lipschitz mappings into hilbert space. *Contemporary mathematics*, 26:189–206, 1984.

Kalajdzievski, D. Scaling laws for forgetting when fine-tuning large language models. *ArXiv preprint*, 2024.

Kleiman, A., Frankle, J., Kakade, S. M., and Paul, M. Predicting task forgetting in large language models. In *ICML 2023 Workshop DeployableGenerativeAI homepage*, 2023.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning, ICML*, 2017.

Kotha, S., Springer, J. M., and Raghunathan, A. Understanding catastrophic forgetting in language models via implicit inference. In *The Twelfth International Conference on Learning Representations*, 2024.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J. N., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2020, 2019.

Lesort, T., Ostapenko, O., Rodríguez, P., Misra, D., Arefin, M. R., Charlin, L., and Rish, I. Challenging common assumptions about catastrophic forgetting and knowledge accumulation. In *Conference on Lifelong Learning Agents*, pp. 43–65. PMLR, 2023.

Li, H., Ding, L., Fang, M., and Tao, D. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308, 2024a.

Li, J., Fang, A., Smyrnis, G., Ivgi, M., Jordan, M., Gadre, S. Y., Bansal, H., Guha, E. K., Keh, S. S., Arora, K., Garg, S., Xin, R., Muennighoff, N., Heckel, R., Mercat, J.-P., Chen, M., Gururangan, S., Wortsman, M., Albalak, A., Bitton, Y., Nezhurina, M., Abbas, A., Hsieh, C.-Y., Ghosh, D., Gardner, J., Kilian, M., Zhang, H., Shao, R., Pratt, S., Sanyal, S., Ilharco, G., Daras, G., Marathe, K., Gokaslan, A., Zhang, J., Chandu, K., Nguyen, T., Vasiljevic, I., Kakade, S. M., Song, S., Sanghavi, S., Faghri, F., Oh, S., Zettlemoyer, L. S., Lo, K., El-Nouby, A., Pouransari, H., Toshev, A., Wang, S., Groeneveld, D., Soldani, L., Koh, P. W., Jitsev, J., Kollar, T., Dimakis, A. G., Carmon, Y., Dave, A., Schmidt, L., and Shankar, V. Datacomp-lm: In search of the next generation of training sets for language models. *ArXiv*, abs/2406.11794, 2024b.

Lin, S., Hilton, J., and Evans, O. 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. doi: 10.18653/v1/2022.acl-long.229.

Liu, Q., Zheng, X., Muennighoff, N., Zeng, G., Dou, L., Pang, T., Jiang, J., and Lin, M. Regmix: Data mixture as regression for language model pre-training. *ArXiv*, abs/2407.01492, 2024.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. The flan collection: Designing data and methods for effective instruction tuning. *arXiv preprint arXiv:2301.13688*, 2023.

Lopez-Paz, D. and Ranzato, M. Gradient episodic memory for continual learning. In *Neural Information Processing Systems*, 2017.

Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., and Hooker, S. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*, 2023.

McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Mehta, S. V., Patil, D., Chandar, S., and Strubell, E. An empirical investigation of the role of pre-training in lifelong learning. *J. Mach. Learn. Res.*, 24:214:1–214:50, 2021.

Mirzadeh, S., Chaudhry, A., Yin, D., Hu, H., Pascanu, R., Görür, D., and Farajtabar, M. Wide neural networks forget less catastrophically. In *International Conference on Machine Learning, ICML*, 2022.

Nguyen, T. T., Huynh, T. T., Ren, Z., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Ostapenko, O., Lesort, T., Rodr'iguez, P., Arefin, M. R., Douillard, A., Rish, I., and Charlin, L. Continual learning with foundation models: An empirical study of latent replay. In *CoLLAs*, 2022.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. Trak: Attributing model behavior at scale. In *International Conference on Machine Learning*, 2023.

Raffel, C. Building machine learning models like open source software. *Communications of the ACM*, 66:38 – 40, 2023.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*, 2016.

Ramasesh, V. V., Lewkowycz, A., and Dyer, E. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*, 2022.

Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., and Almahairi, A. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Sarwar, B. M., Karypis, G., Konstan, J. A., and Riedl, J. Item-based collaborative filtering recommendation algorithms. In Shen, V. Y., Saito, N., Lyu, M. R., and Zurko, M. E. (eds.), *Proceedings of the Tenth International World Wide Web Conference, WWW*, 2001. doi: 10.1145/371920.372071.

Schram, V., Beck, D., and Cohn, T. 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.

Scialom, T., Chakrabarty, T., and Muresan, S. 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.

Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., and Wang, H. Continual learning of large language models: A comprehensive survey. *ArXiv preprint*, abs/2404.16789, 2024.

Smith, J., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R. S., and Kira, Z. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11909–11919, 2022.

Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K. R., Dumas, J., Elazar, Y., Hofmann, V., Jha, A., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J. D., Muennighoff, N., Naik, A., Nam, C., Peters, M. E., Ravichander, A., Richardson, K., Shen, Z., Strubell, E., Subramani, N., Tafjord, O., Walsh, P., Zettlemoyer, L., Smith, N. A., Hajishirzi, H., Beltagy, I., Groeneveld, D., Dodge, J., and Lo, K. Dolma: an open corpus of three trillion tokens for language model pretraining research. *ArXiv preprint*, abs/2402.00159, 2024.

Sun, F.-K., Ho, C.-H., and yi Lee, H. Lamol: Language modeling for lifelong language learning. In *International Conference on Learning Representations*, 2019.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., , and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. *ArXiv preprint*, abs/2210.09261, 2022.

Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. 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. doi: 10.18653/v1/2020.emnlp-main.746.

Tao, M., Feng, Y., and Zhao, D. Can BERT refrain from forgetting on sequential tasks? a probing study. In *The Eleventh International Conference on Learning Representations*, 2023.

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. *ArXiv preprint*, abs/2205.10770, 2022.

Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y., and Gordon, G. J. 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.

Wang, Y., Liu, Y., Shi, C., Li, H., Chen, C., Lu, H., and Yang, Y. 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.

Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.-Y., Ren, X., Su, G., Perot, V., Dy, J., et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648. Springer, 2022.

Wang, Z., Yang, E., Shen, L., and Huang, H. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47:1464–1483, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Winata, G., Xie, L., Radhakrishnan, K., Wu, S., Jin, X., Cheng, P., Kulkarni, M., and Preoţiuc-Pietro, D. Overcoming catastrophic forgetting in massively multilingual continual learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Wu, T., Luo, L., Li, Y.-F., Pan, S., Vu, T.-T., and Haffari, G. Continual learning for large language models: A survey. *ArXiv preprint*, abs/2402.01364, 2024.

Xia, M., Anastasopoulos, A., Xu, R., Yang, Y., and Neubig, G. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020. doi: 10.18653/v1/2020.acl-main.764.

Xia, M., Malladi, S., Gururangan, S., Arora, S., and Chen, D. Less: Selecting influential data for targeted instruction tuning. *ArXiv preprint*, abs/2402.04333, 2024.

Ye, Q., Fu, H. Y., Ren, X., and Jia, R. How predictable are large language model capabilities? a case study on big-bench. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Zhang, X. and Wu, J. Dissecting learning and forgetting in language model finetuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhao, H., Zhou, T., Long, G., Jiang, J., and Zhang, C. Does continual learning equally forget all parameters? In *International Conference on Machine Learning*, 2023.

Zheng, J., Cai, X., Qiu, S., and Ma, Q. Spurious forgetting in continual learning of language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

A. Limitations

We summarize limitations that can be taken as future works.

Forgetting under sequential task learning. In this paper, we analyze and predict forgetting of models that are separately fine-tuned over diverse downstream tasks. We did not explore forgetting under continual or sequential task learning. Although we consider time series forecasting to be relevant to this generalized forgetting prediction problem, we decide to leave it as future work.

Theoretical analysis of low-rank example associations. While we conclude that low-rank example associations are prevalent in LLM forgetting from empirical results, we do not have a theoretical guarantee on why example associations must be low rank, and in which cases the example associations become more complicated. We leave a theoretical understanding of low-rank example associations as future work.

Scope of the experiment setups. Our analysis is restricted to supervised fine-tuning of LMs; other learning setups such as reinforcement learning is not studied. Besides, although we try to cover a broad category of newly learned tasks, the coverage may not be extensive. Due to limitations in computational resources, it was also not possible to repeat experiments under extensive hyperparameter configurations (*e.g.*, learning rate schedulers), which we leave for future work. Nevertheless, we provide an analysis of more learning rate and batch size setups in Appendix I.

Requirement of open access to upstream data. Our analysis requires open access to upstream data to evaluate their forgetting. As a result, our analysis is limited to open-data LLMs. In addition, similar to almost every replay-based algorithm, our forgetting mitigation algorithm requires access to upstream data. We hope that our work can inspire researchers that build closed-data LLMs with in-house data.

Scale of the upstream examples. Due to computational resource restrictions, we significantly subsampled upstream pretraining corpora of LLMs in our analysis. Although expanding the set of upstream examples would not theoretically affect the low-rank observations, we decide to leave larger scale analysis as future works. Besides, we consider that matrix completion based forgetting prediction permits sparse association matrices, where upstream example forgetting z_{ij} are sparsely collected, which we will explore in future work.

Separating out intended forgetting from unintended forgetting. Aligning with most prior works on continual learning, our experiments focus on mitigation of forgetting. We consider the goal aligns with our experiment setup, as upstream examples from Dolma and Tulu are carefully curated and cleaned, and the newly learned tasks intend to accumulate knowledge instead of overwrite previous knowledge. However, we highlight that forgetting is not always detrimental; in fact, the models are often expected to forget (unlearn) sensitive, noisy, and outdated examples (Nguyen et al., 2022). We believe that our analysis of forgetting is helpful for future work that performs targeted forgetting of knowledge.

B. Associations between Learned Tasks and Forgotten Examples

In this section, we detail our analysis about the associations between learned tasks and forgotten upstream examples represented in the $M \times N$ association matrices Z. We visualize the association matrix Z collected in the setups described earlier in Sec. 2. We formally define low-rank approximations and set up quantitative metrics of the complexity of the associations in Sec. B.1, and examine the results of approximation across model types and sizes in Sec. B.2. Lastly, we try to interpret the extracted associations in Sec. B.3.

B.1. Methods and Metrics of Approximating Example Associations

To examine whether simple patterns are dominant in the example associations represented by Z, we attempt to approximate Z with low rank matrices. When Z represents the increase in log perplexity (also known as the loss of language modeling), we fit matrix factorization models $Z^r = \sum_{k=1}^{r} \boldsymbol{\alpha}_k \boldsymbol{\beta}_k^T$ that minimize the Frobenius norm $||Z - Z^r||_F$, where $\boldsymbol{\alpha}_k \in \mathbb{R}^M$, $\boldsymbol{\beta}_k \in \mathbb{R}^N$, r is the rank of the matrix decomposition and k is the index of the component. For binary forgetting measured with exact match drop, we fit a logistic matrix factorization model $Z^r = \sigma(\sum_{k=1}^{r} \boldsymbol{\alpha}_k \boldsymbol{\beta}_k^T)$ that minimizes the cross entropy between Z^r and Z, where σ is the sigmoid function. We measure R^2 or F1 scores of the approximation as the goodness-of-fit metrics.

Interpretations. When $r = 1$, the approximation effectively assumes that the forgetting z_{ij} (or its logit) is a scalar product of a parameter $\beta^{(j)}$ specific to each upstream example and each newly learned task $\alpha^{(i)}$. The set of more forgotten upstream examples is independent of which task the model learns, as $\beta^{(j)}$ trivially determines how fast forgetting uniformly happens on all upstream examples. With a higher rank r, the approximation captures task-dependent forgetting where certain upstream examples are disproportionally more forgotten when learning certain tasks. This inner product formulation is also connected to first-order approximations of loss increase as an inner product of weight updates and gradients (Lopez-Paz & Ranzato, 2017; Lee et al., 2019), where r is the number of LLM parameters.

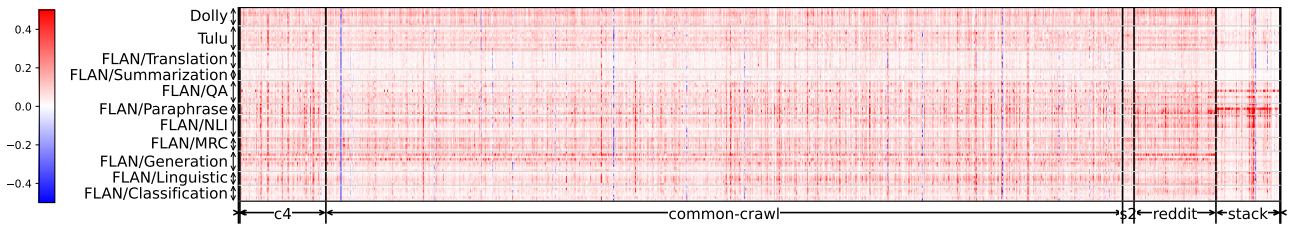
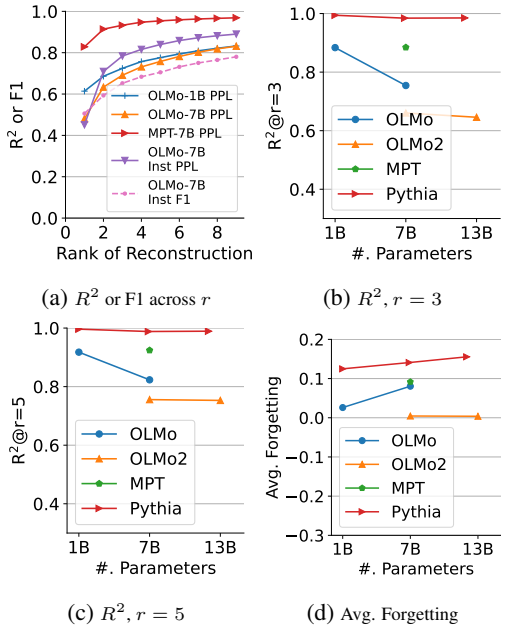


Figure 5. An example of visualized association matrix Z of forgetting between $M = 85$ learned tasks and $N = 141,876$ upstream examples (from Dolma) for OLMo-7B. Each pixel z_{ij} indicates forgetting (in log-perplexity increase) that occurs on an upstream example x_j (in x-axis) after learning a new task T_i (in y-axis). We annotate the domains (*e.g.,* reddit) of upstream examples in the x-axis and the category of each learned task (*e.g.,* FLAN/QA) in the y-axis. We include visualizations of more models and setups in Figure 7 and Figure 9 in Appendix.



(a) R^2 or F1 across r (b) $R^2, r = 3$

(c) $R^2, r = 5$ (d) Avg. Forgetting

Figure 6. (a) R^2 or F1 of the low-rank approximations as we progressively increase the rank of the reconstruction. Forgetting is measured with log perplexity increase or exact match (EM) drop. In (b) and (c), we compare R^2 of approximations at a given rank r across models of different types and sizes over the fixed $M = 19$ tasks from Tulu and Dolly. We also report average upstream example forgetting under various setups in (d) as a reference.

B.2. Examining Complexity of Example Associations

General findings. We present R^2 or F1 of fitting the association with Z^r with progressively higher rank r in Fig. 6(a). We notice that across all training setups of OLMo models, R^2 quickly increases to 0.69 with $r = 3$. Notably, even the rank-1 approximation Z^1 can achieve R^2 scores higher than 0.5. The results suggest that simple patterns are dominant in the associations between learned tasks and forgotten examples.

Example associations across model types and sizes. We compare R^2 of the approximations with a fixed M and N over Pythia, MPT, OLMo, and OLMo2 models. Fig. 6(b) and (c) summarize R^2 at a given rank r. We notice that (1) the goodness of approximations differs among model types. On Pythia and MPT, the R^2 scores at Z^3 are higher than 0.88, while on OLMo-7B and OLMo2, the R^2 scores are around 0.75 and 0.65. (2) The size of the models within the same model family also has an impact on R^2. On Pythia and OLMo2, R^2 stays stable with a slight decrease as the model size increases from 1B to 13B. On OLMo, R^2 is noticeably lower on 7B models compared to 1B. (3) Model families that forget more (*e.g.,* Pythia) tend to produce simpler example associations (higher R^2). However, within the same model family, OLMo-7B results in a lower R^2 score than OLMo-1B despite the fact that the average forgetting is higher.

To summarize, we empirically notice that the associations between learned tasks and forgotten upstream examples are more complicated in more recent and capable LLMs, requiring higher-rank approximations. The complexity of the associations stays stable or increases with larger models within the same family. In Appendix K, we provide more intuitions about how model capability and sizes affect the complexity of the associations with a set of synthetic experiments over MNIST and multi-layer MLPs. In Appendix I,we further study the effect of model training configurations (*e.g.*, learning rate, batch size) on the complexity of the associations Z. We consider that low-rank approximation are prevalent across setups.

Table 2. Correlations between various measures of similarity and upstream example forgetting.

	Pearson ρ	Spearman ρ
Textual (TF-IDF)	-0.049	-0.035
Textual (Representation)	0.021	0.017
⟨Gradient, Weight differences⟩	-0.003	-0.009
⟨Gradient, Gradient⟩	0.061	0.052

B.3. Interpreting Example Associations

Patterns of forgetting from matrix factorizations. The matrix factorization of Z yields interpretable patterns of forgetting in each of its components $\boldsymbol{\alpha}_k \boldsymbol{\beta}_k^T$. As an example, Fig 11 in Appendix visualizes patterns captured by the k-th component in OLMo-7B experiments. We notice the patterns interesting, yet semantically intriguing. For example, on OLMo-7B, Stackoverflow documents are less forgotten when learning summarization tasks, while more forgotten when learning certain paraphrasing tasks.

Correlations between example associations and similarity measures. Are the associations between learned tasks and forgotten examples interpretable from the similarity between the learned tasks and upstream examples? We consider (1) heuristic similarity measures, such as token or representational similarity, and (2) first-order approximations, such as inner products of gradients and inner products between weight updates and gradients (Lee et al., 2019; Doan et al., 2020). We detail each similarity measure in Appendix F. We then evaluate the correlations between the actual forgetting z_{ij} and the various similarity measures on OLMo-1B and summarize the results in Table 2.

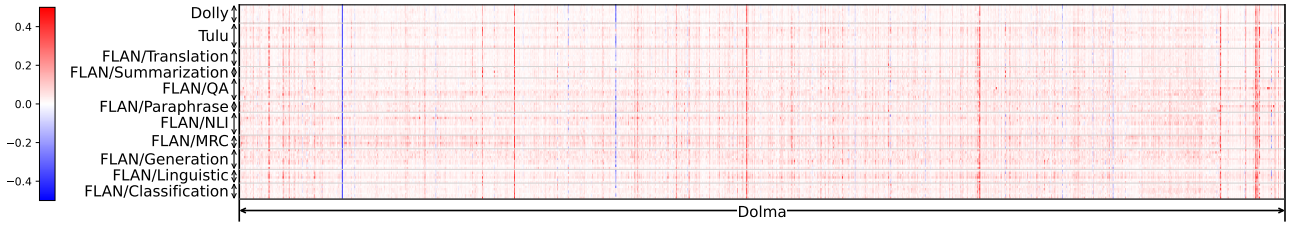
Forgetting correlates poorly with similarity measures of learned tasks and upstream examples. In Table 2, we notice that none of the similarity measures correlates strongly with actual forgetting, with a correlation $|\rho| < 0.1$. These results imply that although simple statistical patterns are dominant in the example forgetting, such associations are not well interpreted with common similarity measures of learned tasks and forgotten examples. Therefore, we hypothesize that leveraging the statistics of forgetting allows better prediction of forgetting than the contents of the tasks and examples, which elicits our next research question about predicting example forgetting.

C. Related Works

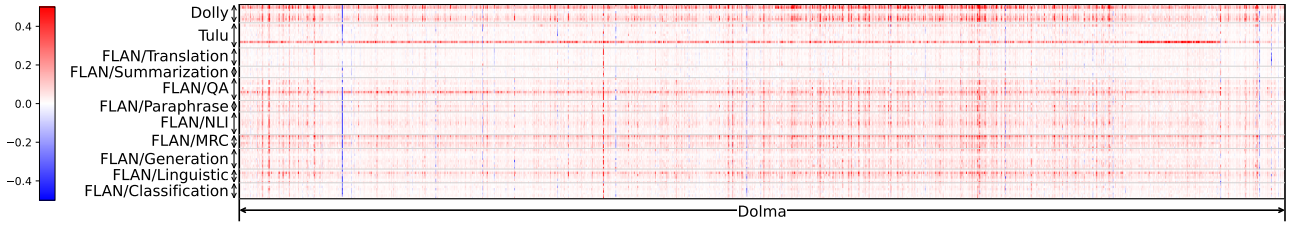
Factors that affect forgetting. In this paper, we primarily studied how the associations between learned and forgotten examples inform forgetting. Prior works have studied various factors that affect forgetting of the models, such as (1) type and size of the LM (Mehta et al., 2021; Scialom et al., 2022; Kalajdzievski, 2024; Mirzadeh et al., 2022; Ramasesh et al., 2022) (2) trainable parts of the model (*e.g.*, LoRA, soft prompts, or full-model tuning) (Biderman et al., 2024a; Razdaibiedina et al., 2023) (3) hyperparameters such as learning rate (Ibrahim et al., 2024; Winata et al., 2023), dropout (Goodfellow et al., 2014), number of training steps (Biderman et al., 2024b; Kleiman et al., 2023) (4) optimizer (Lesort et al., 2023) and training algorithms (*e.g.*, various continual learning algorithms) (Smith et al., 2022; Wang et al., 2022; Shi et al., 2024; Wu et al., 2024), (5) the upstream examples or the knowledge themselves (Toneva et al., 2019; Zhang & Wu, 2024). Future works can study how these factors affect the predictability of forgetting. We consider empirical and theoretical study on the effect of task similarity on forgetting to be most relevant to ours. Ostapenko et al. (2022) empirically study relationships between task similarity and forgetting in foundation models over a sequence of newly learned tasks; our work instead focuses on forgetting of upstream data of LLMs. Theoretical study by Doan et al. (2020); Ding et al. (2024); Evron et al. (2022) dissects effects of the learned tasks on forgetting in linear models or around model initialization. We believe empirical study (Wang et al., 2023; Li et al., 2024a; Zheng et al., 2025) and interpretations of forgetting (Tao et al., 2023; Zhao et al., 2023; Kotha et al., 2024) are 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.

Data selection and data attribution. Related to our work, data attribution studies faithful algorithms to find training examples that account for a prediction (Koh & Liang, 2017; Ilyas et al., 2022) from a pool of training examples. Park et al. (2023); Xia et al. (2024); Li et al. (2024b); Liu et al. (2024) study the problem of selecting a subset of training data that maximizes performance on a given domain or task at a fixed budget for LLMs. Feldman & Zhang (2020); Tirumala et al. (2022); Biderman et al. (2024b); Swayamdipta et al. (2020) identify memorized, important, or forgetful training data. However, the notion of data importance in these works is different from how likely the upstream examples will be forgotten during fine-tuning. Furthermore, a systematical study on how such importance is dependent on newly learned tasks is still absent. Prior works represented by Aljundi et al. (2019a); Wang et al. (2024); Aljundi et al. (2019b); Huang et al. (2024); Sun et al. (2019) study example selection or synthetic example generation strategies for replay-based continual learning algorithms.

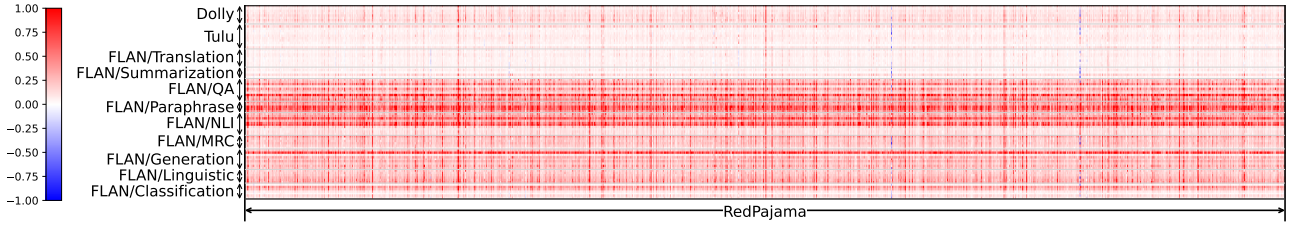
Predicting model behaviors. LLMs can display a hybrid pattern of unpredictable to highly predictable behaviors (Ganguli et al., 2022; Wei et al., 2022). Ye et al. (2023); Xia et al. (2020); Schram et al. (2023) study prediction of task performance across datasets and training setups. We perform prediction at the example level which is more fine-grained and under-explored.



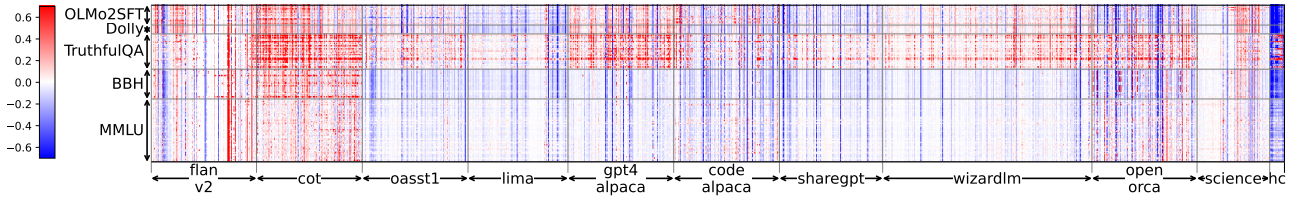
(a) OLMo-1B; forgetting over Dolma after full-parameter fine-tuning on FLAN and Tulu.



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



(c) MPT-7B; forgetting over Dolma after full-parameter fine-tuning on FLAN and Tulu.



(d) OLMo-7B-Instruct; forgetting over Tulu after full-parameter fine-tuning on unseen instruction-tuning tasks.

Figure 7. Additional visualized matrices of associations 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, which indicates no forgetting.

D. Dataset, Model, and LM Training Details

Subsample of upstream datasets. For OLMo experiments, we sample 141,876 text chunks with length 2,048 from Dolma v1.6-sample as upstream examples. For OLMo-7B-Instruct, 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. For OLMo2, we sample 70,000 text chunks with length 2,048 from OLMo2-Mix. For MPT and Pythia, we sample 10,000 2,048-token text chunks from RedPajama and the Pile respectively.

Learned new tasks and their categorization. We summarize the list and the categorization of newly learned tasks in Tables 10 and 11. We also annotate tasks used as in-domain training or test tasks.

Training and evaluation details. For full-parameter fine-tuning of non-instruction-tuned LLMs of all types, we train the model for 1,000 steps with an effective batch size of 8 and a linearly decaying learning rate of $2e^{-6}$. The learning rate is chosen among $\{1e^{-6}, 2e^{-6}, 5e^{-6}\}$ that achieves best average validation perplexity after fine-tuning OLMo-7B on 5 randomly chosen tasks from FLAN. For OLMo-7B-Instruct and MMLU, BBH, TruthfulQA and Dolly, considering the small size of the training sets, we train the models only for 100 steps with an effective batch size of 8. For OLMo2-SFT-Mix tasks, we train the model for 1,000 steps. We use HuggingFace Transformers library for training and VLLM library for efficient inference. Due to the computational resource limitations, the statistics of forgetting are collected in a single run.

Dataset and licenses. MMLU, BBH, and the Pile are released under MIT license. Truthful QA, Dolma, Redpajama, OLMo models, OLMo2 models, Pythia models, and MPT models are released under Apache 2.0 license. Tulu V2, OLMo2-Mix, and OLMo2-SFT-Mix are released under ODC-By license. Dolly is released under CC BY-SA 3.0 license.

Computational Infrastructure. We used 4 Quadro RTX A6000 GPUs for fine-tuning LLMs, and used 1 Quadrio RTX A6000 GPU for LLM inference.

E. Details of Forgetting Prediction and Replay

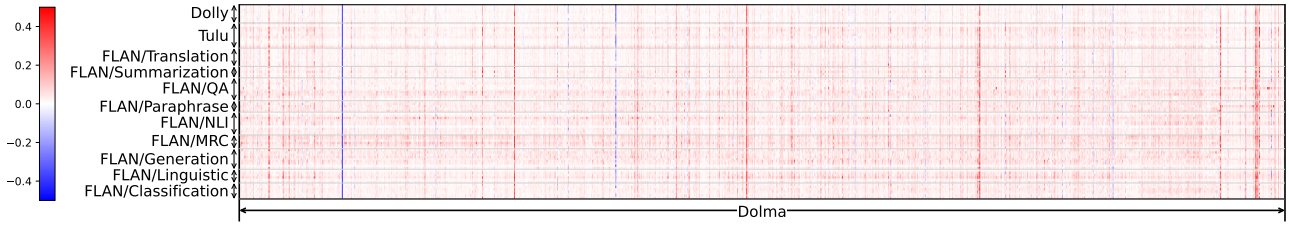
Forgetting prediction algorithms. The additive linear model approximates forgetting as additive effects of learned tasks $\alpha^{(i)}$ and upstream examples $\beta^{(j)}$ ($\alpha^{(i)} + \beta^{(j)}$). The MF models are introduced earlier in Sec. B. Given the seed forgetting $z_i^{\mathcal{S}}$ of a task $T_i \in \mathcal{T}_{\text{test}}$, KNN finds tasks from $\mathcal{T}_{\text{train}}$ that have similar patterns of forgetting over the seed upstream examples $\mathcal{S}$. KNN computes an average of forgetting of top-k similar tasks from $\mathcal{T}_{\text{train}}$ weighted by their similarity as the prediction of forgetting caused by $T_i \in \mathcal{T}_{\text{test}}$ on the upstream examples $x_{1..M}$.

We used all-distilroberta-v1 sentence transformer model as the encoder in the implementation of the embedding similarity based prediction model of forgetting. At inference, given an upstream example x_j, we compute the averaged dot-product with all examples in the learned task T_i. We note that at inference time the approach 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 the approach with ground truth forgetting on these examples.

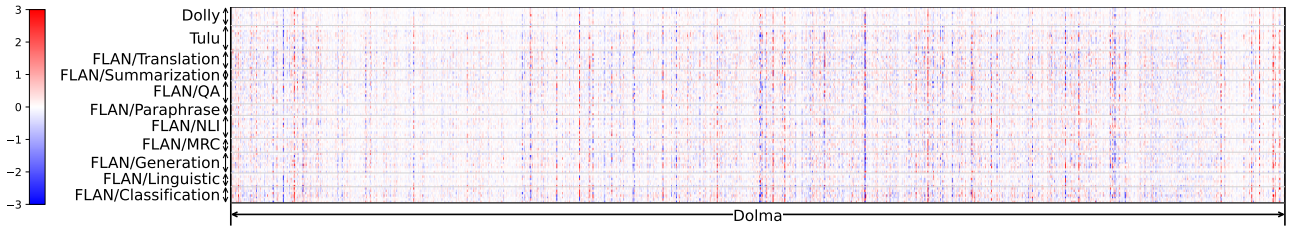
Data Splits for Predicting Example Forgetting. We mark the tasks used as in-domain test splits for predicting example forgetting (Sec. 3) in Tables 10 and 11. The train-test split for the in-domain tasks is randomly generated.

Replaying upstream examples in fine-tuning. We sparsely replay 1 mini-batch of 8 upstream examples every 32 steps of model updates while fine-tuning on new tasks. An exception is fine-tuning of OLMo-7B-Instruct models on Dolly, where we perform a replay every 8 steps given the smaller number of model training steps. Given predicted or ground truth forgetting $z_{i,1..N}$ on upstream examples $x_{1..N}$ 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 while reweighting replay examples with the ground truth forgetting Z.

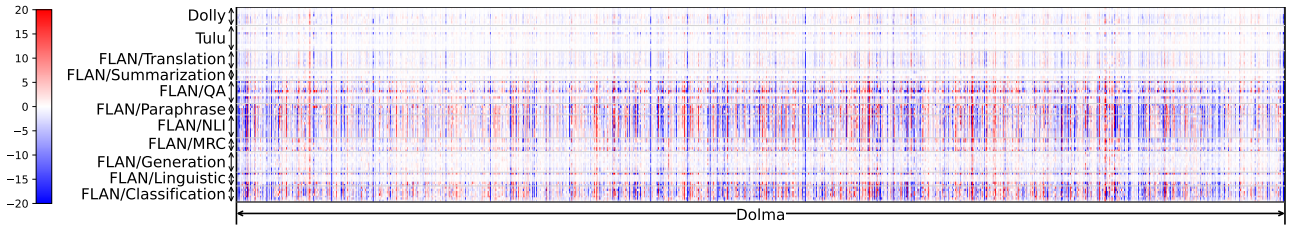
Baselines of replay example selection. We compare with a variant of Maximally Interfered Retrieval (MIR-T) (Aljundi et al., 2019a), a selection strategy sharing the similar notion of importance that forgotten examples should be selected for replay. The approach performs bi-level sampling by selecting the most forgotten examples from a small random subset of upstream data. We extend the approach to select forgotten examples after a full training run on a task, instead of single steps, which achieves better performance. In addition, we apply strategies that consider different definitions of upstream example importance. We examine a strategy based on perplexity thresholds (PPL) (Marion et al., 2023), which samples upstream data of which the perplexity is around the median of the distribution. For OLMo-1B, we also sample replayed examples proportional to the gradient inner products (Grad-Prod) (whose correlation with forgetting is evaluated in Table 2 in Sec. B.3), a representative coreset selection approach that utilizes gradient information (Park et al., 2023; Xia et al., 2024). As a reference, we also experiment with upweighting upstream examples with ground truth forgetting z_{ij}, which, however, is highly inefficient and repetitive to obtain in practice.



(a) Log perplexity increase (forgetting)



(b) Inner products of gradients and weight differences ($z_{ij}^{\text{g-w}}$)



(c) Negative inner products of gradients ($z_{ij}^{\text{g-g}}$)

Figure 8. A side-by-side comparison between the matrices of forgetting, inner products of gradients and weight differences ($z_{ij}^{\text{g-w}}$), and the negative inner products of gradients ($z_{ij}^{\text{g-g}}$) we examined in Sec. B.3.

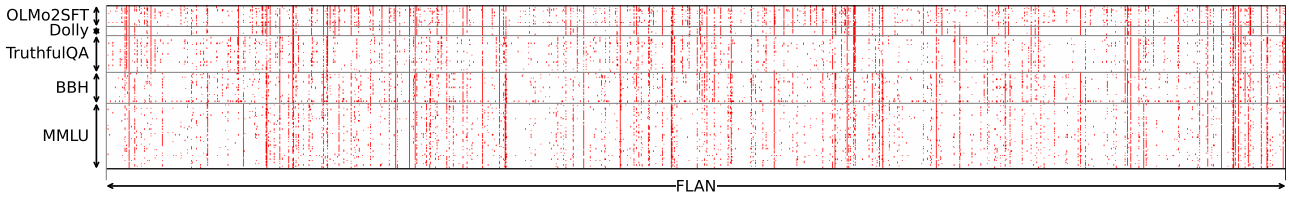
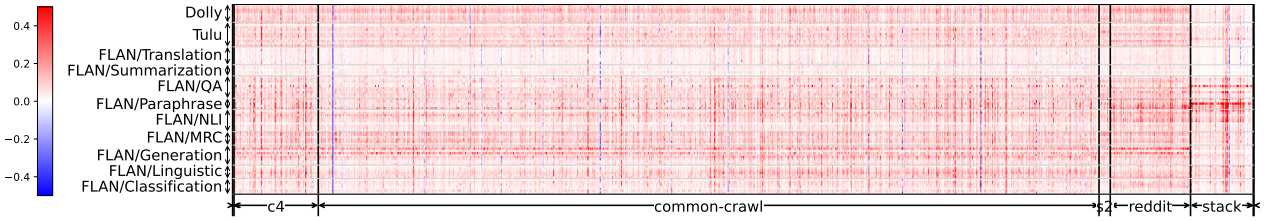
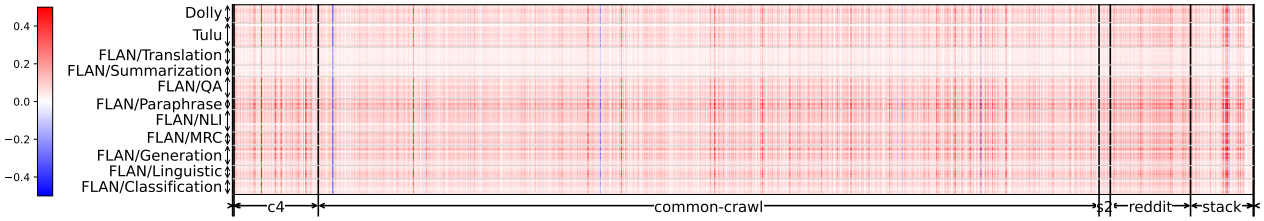


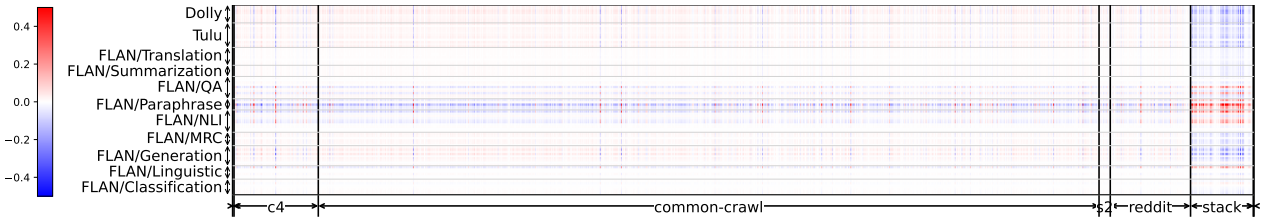
Figure 9. Visualized matrices of associations between learned tasks and forgotten examples on FLAN after fine-tuning OLMo-7B-Instruct, measured with binary exact match drop. Each colored pixel ($z_{ij} = 1$) indicates forgetting of an upstream example x_j after fine-tuning the model on the task T_i.



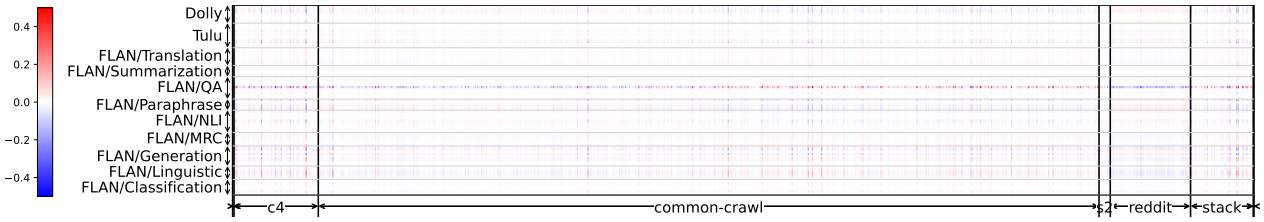
(a) OLMo-7B, full association matrix



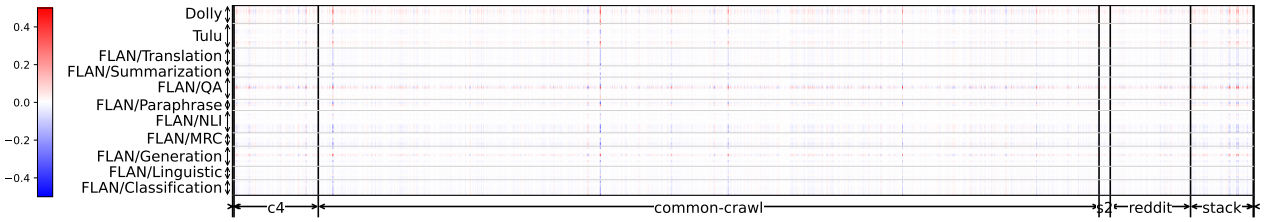
(b) OLMo-7B, compontent $k = 1$



(c) OLMo-7B, component $k = 2$

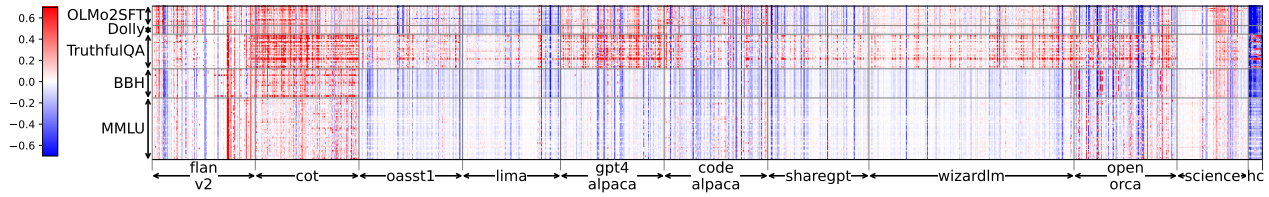


(d) OLMo-7B, component $k = 3$

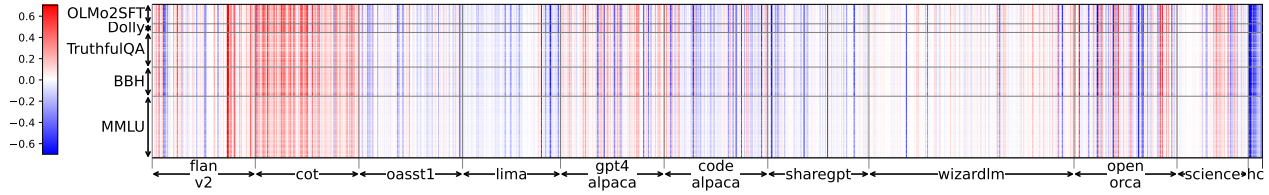


(e) OLMo-7B, component $k = 4$

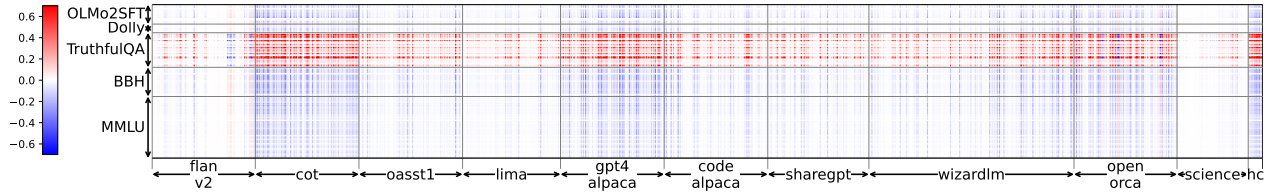
Figure 10. Reconstruction of Z in OLMo-7B experiments with k-th singular value and vectors. Components of higher values of k capture finer-grained details in Z.



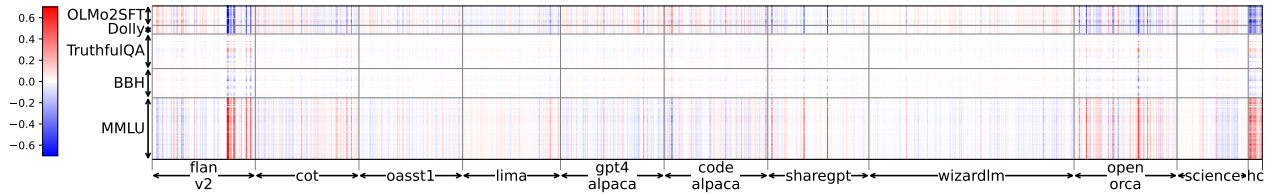
(a) OLMo-7B-Instruct, full association matrix



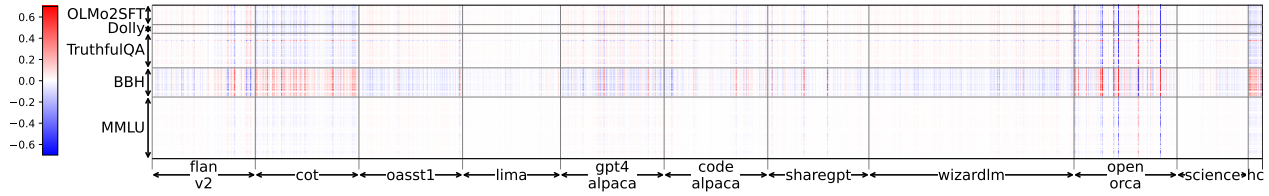
(b) OLMo-7B-Instruct, compontent $k = 1$



(c) OLMo-7B-Instruct, component $k = 2$



(d) OLMo-7B-Instruct, component $k = 3$



(e) OLMo-7B-Instruct, component $k = 4$

Figure 11. Reconstruction of log-perplexity measured forgetting association matrix Z in OLMo-7B-Instruct experiments with k-th singular value and vectors. Higher values of k capture finer-grained details in Z.

Table 3. RMSE of forgetting prediction with matrix factorization (MF) models under different ranks.

Model / Rank	1	3	5	8	10
OLMo-7B	7.31	7.17	7.14	**7.12**	7.14
OLMo-7B-OOD	6.09	5.78	**5.76**	5.77	5.77
OLMo-1B	2.85	2.82	2.80	**2.78**	**2.78**
OLMo-1B-OOD	2.86	2.83	**2.82**	**2.82**	**2.82**

Training and evaluation details. We use Surprise Library 1.1.3[1] for additive linear, MF, and KNN prediction models. For MF, we set the dimension of the learnable features (rank) as 5. We train the regression models for 1,000 epochs over the association matrices.

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

Sensitivity analysis to the rank in MF models. Table 3 summarizes the RMSE of forgetting prediction with MF models across different ranks r. The performance increases with r in the beginning and then drops, indicating an overfit.

F. Details of the Example Similarity Metrics

In this section, we detail the example similarity metrics applied in our analysis in Sec. B.3.

Textual similarity. We calculate the textual cosine similarity between learned tasks and upstream examples using TF-IDF vectorized representations of training examples in each learned task T_i and an upstream example x_j. We also measure text representation similarity with final layer representations of OLMo-1B.

Inner products between projected gradients and model weight updates. The increase of the log perplexity z_{ij} can be approximated with inner products $z_{ij}^{\text{g-w}} = \langle \nabla_\theta f(x_j), \theta_{T_i} - \theta_0 \rangle$ under first-order Taylor expansion (Lee et al., 2019; Doan et al., 2020), where $\nabla_\theta f(x_j)$ is the gradient of the loss of x_j at the initial model before fine-tuning, and $\theta_{T_i} - \theta_0$ are the updates in the model weights after fine-tuning. Following (Park et al., 2023; Xia et al., 2024), we use a random projection matrix $\boldsymbol{P} \sim \mathcal{N}_{|\theta| \times d}(0, 1)$ to reduce the dimension of the gradients or the weight changes to save the cost of storing pre-computed statistics, which preserves the inner products with high probability (Johnson & Lindenstrauss, 1984).

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

Inner products between projected gradients. We also measure the negative inner products of the loss gradients between the upstream example x_j and a learned task T_i, $z_{ij}^{\text{g-g}} = -\langle \nabla_\theta f(x_j), \nabla_\theta f(T_i) \rangle$, as an approximation of forgetting (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019).

G. Effect of Targeted Replay on New-Task Learning and Downstream Performance

Targeted replay does not impede learning of new tasks, as shown with decreased validation perplexity. Table 4 summarizes log perplexity on the validation set of learned tasks after fine-tuning the models with different replay strategies. We notice that example replay and targeted replay often decrease validation perplexity of new tasks. This improvement of new task perplexity can be attributed to less forgetting of general knowledge during fine-tuning. Besides, the results imply that targeted replay does not simply trade off new task learning for reduced forgetting.

LLM leaderboard task evaluation. We evaluate downstream task performance of OLMo-7B-Instruct after fine-tuning on 4 held-out tasks from OLMo2-SFT-Mix (marked in Table 5). We evaluate on Open LLM Leaderboard tasks[2]. For fine-tuned models, we compare no replay, replaying random examples, and replaying with forgetting predicted by offline MF. Tables 5 summarize the results.

We notice that fine-tuning OLMo-7B-Instruct over new tasks without any replay improves metrics scores on some tasks (MUSR, GPQA) but causes forgetting on other tasks (MMLU, BBH, IFEval). Among tasks where performance improved, we do not see benefits of example replay. Among tasks with decreased performance, we see offline-MF mitigates forgetting compared to no replay or random replay slightly.

We conjecture that replay-based approaches are not sufficient to significantly mitigate forgetting on their own, and can be combined with other approaches. We leave more effective algorithms to mitigate downstream task forgetting with predicted forgetting as future works.

H. Towards Interpreting Fine-Grained Associations

We visualize progressive reconstruction with k-th singular value and singular vectors for OLMo experiments in Figure 11.

Semantic meanings of k-th component in the low-rank approximation of the association matrix Z. We perform

[2] https://huggingface.co/docs/leaderboards/open_llm_leaderboard/about

Table 4. Validation log perplexity of the learned tasks after fine-tuning LMs with different replay strategies.

Model	OLMo-1B	OLMo-7B	OLMo-7B	OLMo-7B-Instruct
Dataset	FLAN	FLAN	Tulu & Dolly	OLMo2-SFT-Mix
No Replay	0.9723	0.7736	1.6397	0.6677
Random	0.9621	0.7691	1.6452	**0.6671**
MF-Offline	**0.9601**	**0.7684**	**1.6294**	0.6674

Table 5. Average downstream task performance of OLMo-7B-Instruct models before and after fine-tuning on 4 held-out OLMo2-SFT-Mix tasks.

	MMLU-Pro	BBH	IF-Eval	MUSR	GPQA
Metrics	5-shot Acc	3-shot Acc-norm	0-shot Inst	0-shot Acc-norm	0-shot Acc-norm
Before FT	18.20	37.65	39.93	38.31	26.35
No Replay	17.21	36.76	21.04	**40.20**	**27.99**
Random	17.37	36.85	20.97	39.77	27.73
MF-Offline	**17.43**	**36.89**	**21.14**	39.50	27.69

Table 6. Semantic meaning of k-th component in the factorization of the example association matrix Z. We identify top relevant learned tasks and upstream example domains to k-th component in the factorization of the example association matrix Z.

	OLMo-7B		OLMo-1B	
	Learned Tasks	Forgotten Domain	Learned Tasks	Forgotten Domain
$k = 1$	flan/paws_wiki flan/glue_mrpc flan/story_cloze	None	flan/squad_v2 flan/fix_punct tulu/open_orca	None
$k = 2$	flan/opinion_abstracts_idebate dolly/general_qa flan/story_cloze	StackOverflow (Less)	flan/mnli_matched flan/mnli_mismatched flan/snli	None
$k = 3$	flan/story_cloze flan/fix_punct flan/true_case	None	flan/squad_v2 flan/quac flan/fix_punct	None
$k = 4$	math_dataset dolly/general_qa flan/opinion_abstracts_idebate	None	flan/rte flan/opinion_abstracts_idebate flan/story_cloze	None

Table 7. R^2 of approximating forgetting association matrix Z across different learning rates (LR) and batch sizes (BS). The default learning rate and batch sizes are 2e-6 and 8. We report results at $M = 19$ learned tasks from Tulu and Dolly and $N = 141,871$ upstream examples from Dolma.

Model	Setting	$r = 2$	$r = 3$	$r = 5$
OLMo-7B	Default	0.6981	0.7542	0.8232
	LR=1e-6	0.7071	0.7697	0.8361
	BS=32	0.8041	0.8453	0.9011
	LR=5e-6	0.9331	0.9582	0.9718
	LoRA, LR=1e-4	0.8963	0.9307	0.9562
OLMo-1B	Default	0.8344	0.8836	0.9177
	LR=1e-6	0.8471	0.8989	0.9214
	BS=32	0.8724	0.9289	0.9607

further analysis into the patterns captured by the k-th singular value and singular vectors by identifying the most relevant learned tasks and upstream example domain to the component. For each k and its corresponding component $\boldsymbol{\alpha}_k \boldsymbol{\beta}_k^T$, we extract top 3 rows with the highest mean (*i.e.*, top 3 relevant learned tasks T_i). We also extract top 50 columns with highest mean (*i.e.* top 50 relevant upstream examples) and the domain where these upstream examples are drawn from. For OLMo models, the domains are one of C4, common-crawl, Gutenberg books, Reddit, Science, StackOverFlow, and Wikipedia. We compare the distribution of domains in the top 50 upstream examples, and perform a z-test to determine upstream example domain that is significantly more or less forgotten compared to a prior domain distribution of top 50 most forgotten upstream examples (columns with highest mean in Z). The results are summarized in Table 6.

We highlight some notable patterns in Table 6. (1) Some component Z_k highlights forgetting patterns of upstream examples from certain domains. On OLMo-7B, the second component ($k = 2$, also visualized in Figure 11(c)) highlights patterns where StackOverFlow examples are disproportionally more or less forgotten. (2) Some components Z_k highlight forgetting when learning specific types of tasks. For example, the second component ($k = 2$) on OLMo-1B highlights forgetting patterns after learning NLI tasks (mnli_matched, mnli_mismatched, snli). This also exemplifies how learning similar tasks causes similar sets of upstream examples to be more forgotten.

I. Effect of Model Training Setups on Forgetting

In this section, we examine patterns of forgetting and its low-rank approximations across various training setups.

Table 8. Sample Pearson correlation coefficient between upstream example forgetting under different replay strategies. The results are collected with OLMo-7B models fine-tuned on 8 Dolly tasks.

Setup A	Setup B	Sample Pearson r
No Replay	No Replay + seed change	0.9403
No Replay	Random Replay	0.7919
No Replay	MF-Offline Replay	0.8121
Random Replay	MF-Offline Replay	0.7709

Low-rank approximations hold at various learning rates, batch sizes, and LoRA fine-tuning. Table 7 summarizes R^2 of low-rank approximations. Reducing learning rate from the default 2e-6 to 1e-6 causes very minor change of R^2; Nevertheless, we notice increased R^2 under a larger learning rate like 5e-6, indicating even simpler associations between the learned tasks and forgotten upstream examples. Under this overly large learning rate, the model suffers substantially more forgetting on all examples. Besides, increasing the batch size (under the same number of parameter update steps) causes an increase in R^2, indicating simpler associations, possibly due to reduced variance of learned models under a larger batch size. LoRA fine-tuning (Hu et al., 2022) results in simpler associations despite the model forgets less than full parameter fine-tuning.

Replay can change the patterns of forgetting compared to no replay, but the change is mild. We compute sample Pearson correlation coefficient between upstream example forgetting collected under different replay strategies, and more specifically (1) no replay (2) random replay and (3) MF-Offline replay on OLMo-7B and 8 Dolly tasks. We evaluate forgetting of upstream examples held-out from replay. We include correlations between upstream example forgetting under two different training random seeds as the baseline.

We summarize the results in Table 8. The correlations between no replay and replay are around 0.8, lower than the 0.94 baseline, but are still strongly positive. Replay example selection strategies (random or MF-offline) do not have a strong impact on the correlation. The results imply replay changes patterns of forgetting to a mild degree. Future works can study the limit where the patterns of forgetting experience notable changes.

J. Combining Embedding and Matrix Completion based Prediction

Although our results in Sec. 3 demonstrate subpar performance of the embedding-based approach, we show combining embedding and matrix completion approaches leads to improvement over the both. Specifically, we fit the residual

Table 9. Combining matrix completion and embedding based prediction of forgetting. We report RMSE (↓) or F1(↑) of predicting example forgetting over a held-out set of upstream examples after fine-tuning LMs on unseen new tasks.

	In-Domain					Out-of-Domain				
Model	OLMo-1B	OLMo-7B	MPT	OLMo-7B Instruct		OLMo-1B	OLMo-7B	MPT	OLMo-7B Instruct	
Metrics	RMSE	RMSE	RMSE	RMSE	F1	RMSE	RMSE	RMSE	RMSE	F1
Matrix Completion	2.79	7.14	10.41	13.74	58.16	2.82	5.76	7.03	38.90	43.57
Embedding	2.81	7.44	13.86	13.94	55.46	3.53	6.16	10.59	41.22	42.95
Embedding + MC	**2.76**	**7.09**	**10.20**	**13.25**	**59.36**	**2.76**	**5.75**	**7.00**	**40.36**	**42.98**

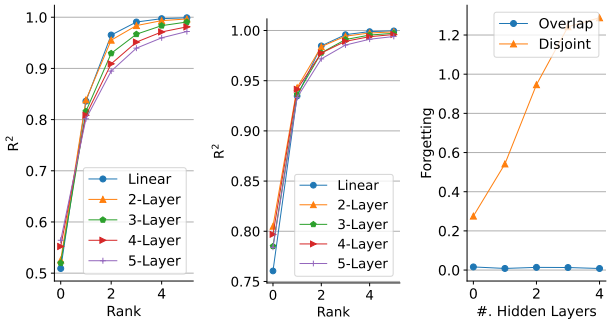
error of the matrix completion model $Z - Z_{MC}$ with the embedding model, where Z_{MC} is the prediction given by the matrix completion algorithms. We use the better of the additive and MF models reported in Table 1. Table 9 summarizes the results. We see that combining the two approaches improves performance over the two.

We consider the embedding based approaches better capture fine-grained knowledge conflicts that leads to forgetting of one example by learning the other (*e.g.*, two documents stating contradictory facts) that is missed by matrix completion approaches that rely on statistics of forgetting.

K. Synthetic Dataset Experiments

In this section, we present a set of experiments on a synthetic dataset, Rotated-MNIST, broadly used in continual learning research. We aim to provide intuition for future research about how the complexity of the example associations can depend on (1) the coverage of knowledge represented in upstream examples, and (2) the size of the models, in highly controlled setups. We apply a training setup that resembles pretraining and fine-tuning paradigm in transformer language models. Specifically, the models are first pre-trained on 10 rotated variants (0-90°) of the MNIST digit classification dataset (as upstream tasks and examples). Then, the models are fine-tuned on one of 40 unseen rotations for one epoch. Among the 40 unseen rotations, 20 are drawn from the same range as upstream examples (0-90°), while the other 20 are drawn from a disjoint range (-90-0°). This separation controls the amount of shared knowledge between the newly learned and upstream tasks.

We mostly apply other training setups and hyperparameters in Aljundi et al. (2019a). Each task (rotation) includes 1,000 training examples. We train a MLP classifier to predict among 10 digits given an input image without providing its rotation (or the task identifier). We experiment MLP classification models with 1 layer (a linear model) to 5 layers. We collect the example associations Z, and visualize them in Figure 13. The upstream and the newly learned rotation tasks are ordered by their rotations in the x or y axis.

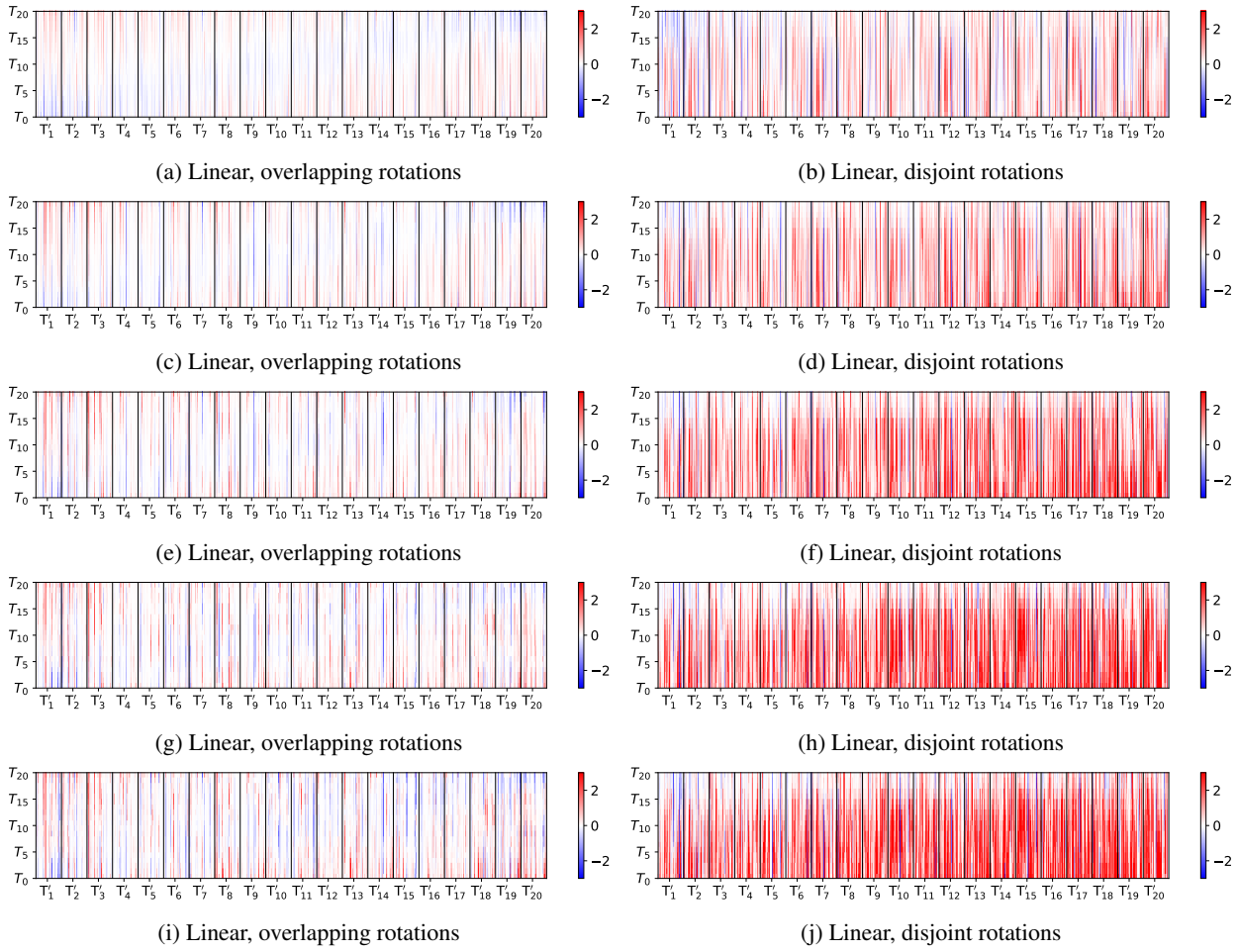


(a) Overlapping rotations (b) Disjoint rotations (c) Avg. Forgetting

Figure 12. R^2 of low-rank approximations of the example association in the forgetting of MLP models on rotated MNIST experiments. We report R^2 when the rotation of the newly learned task overlaps or is disjoint with the upstream data separately. y-axes are not in the same scale.

Effects of knowledge coverage. When the rotations of the newly learned tasks overlaps with the range of upstream examples, the forgetting is harder to approximate with low-rank approximations, resulting in a R^2 only around 0.5 for all MLP models with rank-1 approximation. In contrast, the R^2 scores are much higher when the rotations do not overlap, alongside higher average forgetting; R^2 with rank-1 approximation is higher than 0.8 in these setups. The results imply that the amount of shared knowledge between fine-tuning tasks and upstream examples can have an impact on the complexity of the example associations. In other words, models trained with a broad coverage of upstream examples, or that cover knowledge required for diverse fine-tuning tasks, can yield more complicated example associations. In the context of LLMs, we have noticed that more powerful LLMs (such as OLMo and OLMo2, compared to Pythia and MPT) with broader coverage of knowledge yield more complicated patterns of forgetting.

Effects of model sizes. We compare the MLPs of different number of layers trained on the same upstream data. From Figure 12, the patterns of forgetting are nosier in deeper models. The R^2 scores at rank 3 or 5 decreases with added layers in MLPs, providing a quantitative measure of



(a) Linear, overlapping rotations

(b) Linear, disjoint rotations

(c) Linear, overlapping rotations

(d) Linear, disjoint rotations

(e) Linear, overlapping rotations

(f) Linear, disjoint rotations

(g) Linear, overlapping rotations

(h) Linear, disjoint rotations

(i) Linear, overlapping rotations

(j) Linear, disjoint rotations

Figure 13. Example associations between learned tasks and forgotten upstream examples on Rotated MNIST with overlapping or disjoint ranges of rotations. We measure increase in the cross-entropy loss as forgetting.

increased complexity between learned tasks and forgotten examples.

To summarize, our analysis with synthetic datasets provide intuition about the effect of knowledge coverage and model sizes on the complexity of example associations in forgetting. We hope the set of synthetic experiments can inspire more comprehensive study on how the complexity of example associations are affected by various factors in future works.

L. Broader Impact

Our research tries to better demystify and mitigate forgetting in LLM fine-tuning. We expect that the better understanding alongside the reduced forgetting can, in turn, encourage model developers to promptly update their LLMs to address limitations of the models and improve their models in continuing efforts. We expect the broader application of the continual learning practice will reduce training costs compared to re-training models, and ultimately result in more powerful models under a controlled training budget.

Although we do not see direct negative impact of predicting example forgetting, we highlight that in real-world continual learning setups, blindly mitigating forgetting may result in outdated knowledge and data privacy breaches in LLMs. Dissecting beneficial and intended forgetting from unintended or catastrophic forgetting requires attention in real-world setups.

Task Category	Task	Task Category	Task
FLAN/Classification	aeslc	FLAN/QA	arc_challenge*
	ag_news_subset		arc_easy*
	imdb_reviews		bool_q
	sentiment140		coqa*
	sst2		cosmos_qa
	trec*		math_dataset*
	yelp_polarity_reviews*		natural_questions*
FLAN/Linguistic	cola		openbookqa*
	definite_pronoun_resolution*		piqa
	fix_punct*		trivia_qa*
	true_case	FLAN/Summarization	cnn_dailymail
	word_segment		gigaword
	wsc*		multi_news
FLAN/Generation	common_gen		samsum
	copa		wiki_lingua_english_en
	dart	FLAN/Translation	para_crawl_enes
	e2e_nlg*		wmt14_enfr
	hellaswag		wmt16_translate_csen
	opinion_abstracts_idebate*		wmt16_translate_deen
	opinion_abstracts_rotten_tomatoes		wmt16_translate_fien
	story_cloze		wmt16_translate_roen
	web_nlg_en		wmt16_translate_ruen*
FLAN/MRC	drop		wmt16_translate_tren*
	multirc	Tulu	open_orca
	quac		oasst1
	record		lima
	squad_v1		code_alpaca
	squad_v2		gpt4_alpaca
FLAN/NLI	anli_r1		cot
	anli_r2		science
	anli_r3		flan_v2
	cb*		sharegpt
	mnli_matched		hard_coded
	mnli_mismatched		wizardlm
	qnli*	Dolly	brainstorming
	rte		closed_qa
	snli		information_extraction
	wnli		classification
FLAN/Paraphrase	glue_mrpc		open_qa
	glue_qqp*		general_qa
	paws_wiki		creative_writing
	stsb		summarization
	wic*		

Table 10. The list of learned tasks in our experiments on OLMo-1B, OLMo-7B and MPT-7B. * notes for tasks used as the in-domain test split in forgetting prediction experiments in Sec. 3.

Task Category	Task	Task Category	Task
MMLU	abstract_algebra		object_counting*
	anatomy		penguins_in_a_table
	astronomy		reasoning_about_colored_objects
	business_ethics		ruin_names
	clinical_knowledge		salient_translation_error_detection
	college_biology*		snarks
	college_chemistry		sports_understanding
	college_computer_science		temporal_sequences
	college_mathematics		tracking_shuffled_objects_five_objects
	college_medicine*		tracking_shuffled_objects_seven_objects
	college_physics		tracking_shuffled_objects_three_objects
	computer_security		web_of_lies
	conceptual_physics*		word_sorting
	econometrics	TruthfulQA	Nutrition
	electrical_engineering		Stereotypes
	elementary_mathematics		Confusion
	formal_logic		Psychology
	global_facts*		Language
	high_school_biology*		Sociology
	high_school_chemistry		Finance
	high_school_computer_science		Indexical Error
	high_school_european_history*		Science
	high_school_geography		Misconceptions
	high_school_government_and_politics		Economics
	high_school_macroeconomics		Education
	high_school_mathematics		Proverbs
	high_school_microeconomics		Conspiracies
	high_school_physics*		Religion
	high_school_psychology		Statistics
	high_school_statistics		Misquotations
	high_school_us_history*		Subjective
	high_school_world_history		Law
	human_aging*		History
	human_sexuality*		Fiction
	international_law		Mandela Effect
	jurisprudence		Politics
	logical_fallacies*		Misinformation
	machine_learning		Logical Falsehood
	management*		Distraction
	marketing*		Weather
	medical_genetics		Myths and Fairytales
	miscellaneous		Superstitions
	moral_disputes		Advertising
	moral_scenarios*		Paranormal
	nutrition		Health
	philosophy*	Dolly	brainstorming
	prehistory		rte closed_qa
	professional_accounting		snli information_extraction
	professional_law		wnli classification
	professional_medicine*		FLAN/Paraphrase glue_mrpc open_qa
	professional_psychology		glue_qqp* general_qa
	public_relations*		paws_wiki creative_writing
	security_studies		stsb summarization
	sociology*	OLMo2SFT-Mix	coconot*
	us_foreign_policy*		evol_codealpaca_heval
	virology		flan_v2
	world_religions		no_robots
BBH	boolean_expressions*		numinamath_tir_math*
	causal_judgement		oasst1
	date_understanding		personahub_code
	disambiguation_qa		personahub_ifdata
	dyck_languages*		personahub_math*
	formal_fallacies*		aya
	geometric_shapes		open_math_2_gsm8k
	hyperbaton*		personahub_math_interm_algebra
	logical_deduction_five_objects*		sciriff
	logical_deduction_seven_objects		synthetic_finalresp_wildguardmixtrain
	logical_deduction_three_objects		table_gpt
	movie_recommendation*		wildchat
	multistep_arithmetic_two		wildjailbreak*
	navigate		personas-math-grade

Table 11. The list of learned tasks in our experiments on OLMo-7B-Instruct. * notes for tasks used as the in-domain test split in forgetting prediction experiments in Sec. 3.