Gradient Localization Improves Lifelong Pretraining of Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) trained on web-scale text corpora have been shown to capture world knowledge in their parameters. However, the mechanism by which language models store different types of knowledge is poorly understood. In this work, we examine two types of knowledge relating to temporally sensitive entities and demonstrate that each type is localized to different sets of parameters within the LLMs. We hypothesize that the lack of consideration of the locality of knowledge in existing continual learning methods is responsible for failed uptake of new information and catastrophic forgetting of previously learned information. We demonstrate that targeted training to these relevant layers can improve the performance of continually learned language under temporal drift.

1 Introduction

Pretraining over diverse datasets has been shown to encode world knowledge in the parameters of large language models (LLMs) (Petroni et al., 2019; Roberts et al., 2020; Gueta et al., 2023) from massive static web-scale datasets. However, these models are frequently trained on large static text corpora which are unable to reflect changes in world knowledge or language usage that occur after the initial data collection. In practice language models are deployed in dynamic real-world settings, and their learned knowledge becomes stale over time; the temporal degradation can be evaluated according to intrinsic measures such as perplexity, or extrinsic downstream performance (e.g. question answering) (Lazaridou et al., 2021; Luu et al., 2022; Dhingra et al., 2022; Yao et al., 2022; Nylund et al., 2023; Cheang et al., 2023).

Incrementally training of language models on streams of data which reflect the changes in language usage and world knowledge has been explored as a method to mitigate temporal performance degradation without incurring the heavy computational and environmental costs of retraining models on large pretraining corpora (Jang et al., 2021, 2022; Lin et al., 2022). However, naive online finetuning on these datastreams has been observed to: induce hallucinations in model generations (Kang et al., 2024), failures to uptake new information (Hu et al., 2023a), and catastrophic forgetting of previously learned information (Zhu et al., 2020). To address these problems, recent work has explored continual learning and online learning methods for adapting large language models on streams of documents (Loureiro et al., 2022; Scialom et al., 2022; Jang et al., 2022)

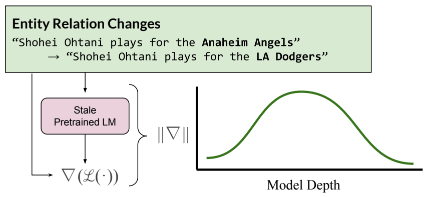

Figure 1: The NLL loss gradients of updated entities and newly mentioned entities observe characteristic patterns of layers with large norms.

As one potential solution, continual pretraining has been shown to improve performance when training on a sequence of natural language domains (Gururangan et al., 2020), but these methods often fail to acquire new knowledge (Hu et al., 2023a; Onoe et al., 2023). While continual learning methods have been shown to mitigate temporal degradation on the task-level, the mechanisms by which neural language models store and update information are not well understood: Appendix C contains details of related work.

In this work, we consider the practical continual language learning setting of temporal language drift and probe the performance of language models on two types of entity relationships known to observe temporal degradation: (1) acquisiton of information about new entities, and (2) updating relationships between existing entities. We hypothesize that the poor performance of existing continual learning methods on these tasks can be in part attributed to a misalignment in the autoregressive language modeling pretraining objective and the ideal parameter updates required to acquire new information or update existing knowledge. As an indicator of this misalignment, we examine models' gradient updates computed on knowledge intensive salient entity spans and compare them with those seen instandard continual pretraining, and discover that the gradient norms observe high values in distinct groups of layers based on the type of entity relationship presented in the sequence (see Fig. 1).

Based on these observations, we propose new methods for aligning the updates steps during continual pretraining to better align with the Through empirical study, we show that the observed characteristic gradient patterns occur across autoregressive, transformer language models of various of sizes; and we demonstrate the efficacy of our proposed method through performance improvements on knowledge probing tasks when applied on top of existing continual learning methods in pretraining.

2 Knowledge Probing with Salient Span Prediction

We probe language models with the problem of salient span prediction, which has previously shown success as a pretraining objective for knowledge-intensive tasks such as closed-book question answering (Cole et al., 2023; Guu et al., 2020). In salient span prediction, a model is provided with a sequence and tasked with completing a masked slot corresponding to a named entity or noun phrase. Specifically, we examine language models on probing tasks for temporal entity knowledge in which the masked sequence corresponds to an update existing of knowledge about temporally sensitive entities or is a mention of an emerging new entities that was not previously seen during pretraining.

2.1 Probing Datasets

We study these using the Dynamic TempLAMA (Dhingra et al., 2022) and the Entity Cloze By Date (ECBD) (Onoe et al., 2022) diagnostic datasets, respectively. Examples can be found in Table 3. The Dynamic TempLAMA dataset contains slot-filling cloze queries where the goal is to complete a subject-object relation in which there are multiple candidate object answers that change over time. Examples are generated from natural language templates based on subject-object relations extracted from Wikipedia metadata, and are generated sequentially for three month periods. For our analysis, we examine splits for each year from 2019 to 2021. As the subject in each example has been mentioned in both the seen and unseen data, we use this dataset to evaluate the ability of continual learning techniques to *update* existing knowledge. To evaluate continual learning methods in knowledge acquisition about new entities, we consider the ECBD dataset which consists of sentences reference emerging entities. Examples consist sentences containing the emerging entity with the goal of predicting noun-phrase spans related to the target entity. Examples are grouped by year, according to the first time of mention.

2.2 Models

We examine decoder-only transformer language models of various sizes, specifically: GPT 2-Base (110M parameters) and GPT-2 Large (770M parameters). To evaluate the perplexity of each of these models, we provide the example context of each example up to the salient span and compute the perplexity over the salient span as in (Onoe et al., 2022, 2023).

To align the each language model with each Wikipedia-based knowledge for the probing tasks, we perform domain adaptive pretraining on snapshots of Wikipedia retrieved prior to the pretraining data cutoffs for each model. We perform initial pretraining GPT-2 models on Wikipedia snapshots from January 2019; GPT-Neo from January 2020.

2.3 Probing Model Response to Salient Spans

We hypothesize that the portions of the model responsible for different forms of knowledge can be identified by tracing the gradient norm of examples which reflect the target form of knowledge.

For the ECBD probing dataset, we examine the loss gradient with respect to the salient span corresponding to the target entity or its related noun phrase, which we refer to as ECBD-ENT and ECBD-NP, respectively. For the TempLAMA dataset, we examine the loss gradient with respect to the object noun phrase.

Beginning with a domain-adapted model pretrained on a snapshot of Wikipedia from 2019, we examine the average per-token loss gradients of the salient spans from the 2019 splits of TempLAMA

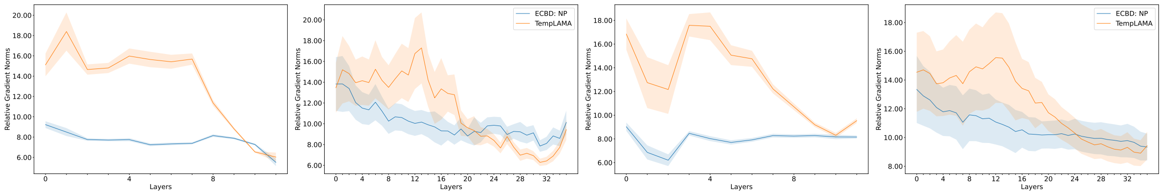

Figure 2: Relative gradient norms for the salient spans in ECBD and TempLAMA for the GPT-2 Base (110M), *(1,3)*, and GPT-2 Large (770M; *(2,4)*), models. Norms for attention (*1,2*) and norms for MLP *(3,4)* are depicted separately. Gradient norms of salient spans are 4 to 15x larger than those of the full sequence.

and ECBD. For comparison, we compute the gradient of the loss for 2000 examples from the 2019 Wikipedia snapshot over the full sequence.

Precisely, we provide the autoregressive language model with the left context preceding the salient span and compute the parameter gradient with respect to the loss averaged over each token in the target span tokens. We then aggregate the gradients according to their respective transformer block, and component attention and MLP layers and compute the L2-Norm of the gradients for each layer.

For the GPT-2 Base model, the gradient norms of each attention and MLP block for the salient spans probes are consistently 4 to 15x higher than the gradient norms of the randomly sampled pretraining examples for all transformer layers. Additionally, we observe that salients spans corresponding to changes in entity relations observe a distinct profile in which they exhibit large magnitude in the early and middle layers and are larger in the attention layers than in the MLP layers.

3 Gradient Localized Continual Pretraining

Ideally, naive pretraining of a language model on a changing stream of data would be sufficient to update a model to capture the relevant changes in knowledge. However, recent work has demonstrated that current methods for continual learning often suffer from both catastrophic forgetting and a failure to uptake new knowledge even when it is directly contained in the training corpus (Hu et al., 2023a; Kang et al., 2024). We hypothesize that failed transfer occurs due to a misalignment of the NLL objective with the information content of the data observed during continual pretraining.

Based on our observations from §2, we hypothesize that the acquisition of entity knowledge can be improved by amplifying updates to the layers are relevant to the learning of salient entity spans. To identify these relevant layers, we compute the relative gradient norm for each layer as the ratio between the gradient norm $\tilde{\nabla}_i$ in the layer i w.r.t. randomly sampled data from the continual pretraining data stream, and data sampled from the validation set of the TempLAMA diagnostic dataset:

$$\frac{||\nabla_i \mathcal{L}(M_\theta, (x,y)_{\text{TempLAMA}})||}{||\nabla_i \mathcal{L}(M_\theta, (x,y)_{\text{PT}})||} \tag{1}$$

We propose two methods for aligning gradient updates during continual pretraining with to improve knowledge uptake by tracing the gradient magnitudes for relevant salient spans from the TempLAMA diagnostic dataset based on the relative gradient norms traced through each layer. We refer to our methods as Traced Gradient Layers (TGL).

Selecting Trainable Layers for Pretraining Based on Relative Gradient Norm We consider a simple approach to target continual pretraining updates to layers with high relative gradient norm, by only updating parameters where the relative gradient norm on the TempLAMA diagnostic dataset exceed the mean relative gradient norm of all layers – we refer to this parameter freezing method as TGL + FP. In the case of the GPT-2 architecture, we separate the model into its component MLP and attention layers, then compute the relative gradient norm for each layer as the ratio between the average gradient norm computed over samples from both the TempLAMA dataset and the continual pretraining corpus. Precsisely, we freeze a parameter group i if $\tilde{\nabla}_i < \frac{1}{\text{No. Layers}}(\sum_{k \in \text{Layers}} \tilde{\nabla}_k)$.

Per-Layer Adaptive Learning Rates from Relative Gradient Norm Rather than using relative gradient norm as a hard threshold to determine which layers to update, we instead consider an adaptive approach in which we set the learning rate for layers to scale with the magnitude of the rela-

Evaluation Set: 2020	ECBD Pop.	ECBD NP	TempLAMA
Pretrain	40.99	47.44	81.92
Domain Pretrain	30.90	41.39	62.99
Continual Pretrain	34.79	**43.97**	56.72
+ TGL with FP	**34.13**	44.20	**55.19**
LoRA: 64D, Attn	31.94	41.40	57.21
+ TGL with FP	**30.28**	**41.05**	**56.32**
MixReview	28.70	**37.34**	67.64
+ TGL with FP	**28.24**	37.77	**60.05**
RecAdam	34.78	43.92	57.34
+ TGL with FP	**33.56**	**43.41**	**54.75**

Table 1: TGL with frozen layers improves performance (perplexity of slot) of GPT2-Large (770M) during continual pretraining.

tive gradient norm. We scale the per-layer learning rate for layer i as : $\eta_i = \eta \frac{\tilde{\nabla}_i}{\max_{i \in \text{Layers}}(\tilde{\nabla}_k)}$

3.1 Baselines

We compare the performance of our proposed continual pretraining method with existing approaches from continual learning. We consider vanilla continual pretraining in which we update all parameters; a prarameter-expansion method LoRA (Hu et al., 2021), which introduces additional trainable low rank adapters to the self-attention layers; a replay-based method MixReview (He et al., 2021), which adds previously seen data is randomly mixed alongside current data during continued pretraining; and a regularization-based method of RecAdam (Chen et al., 2020), which imposes a quadratic penalty on the norm of parameter updates. We provide full details on the training datasets and hyperparameters in the Appendix.

3.2 Evaluating TGL for Continual PT

To evaluate the performance of TGL+FP and TGL+AR, we perform domain adaptive pretraining of GPT-2 Base and Large on the complete Wikipedia corpus from January 2019 for 4 epochs, then incrementally train on the complete set of Wikipedia revisions for the subsequent years of 2020 and 2021. To evaluate the performance of these models, we probe the continually pretrained model after each updating on new year of Wikipedia revisions using the corresponding temporally delineated split from the ECBD-NP and TempLAMA test datasets 2.1. To evaluate whether either TGL method leads to catastrophic forgetting, we also report performance on ECBD-Popular, sequences referring to entities common in all years.

In Table 2, we report the perplexities of the continually pretrained model on the 2020 test splits

Evaluation Set: 2020	ECBD Pop.	ECBD NP	TempLAMA
Pretrain	78.61	80.04	162.54
Domain Pretrain	55.26	62.59	80.51
Continual Pretrain	64.13	72.42	83.39
+ TGL with ALR	**57.62**	**64.83**	77.58
+ TGL with FP	57.75	65.08	**74.55**
MixReview	54.10	61.54	82.16
+ TGL with ALR	53.50	**61.01**	77.04
+ TGL with FP	**53.48**	61.48	**76.35**
LoRA	55.77	**65.56**	80.11
+ TGL with ALR	57.75	69.44	**78.40**
+ TGL with FP	**58.09**	67.62	78.77
RecAdam	57.55	**64.60**	76.67
+ TGL with ALR	**57.52**	64.77	77.32
+ TGL with FP	57.55	64.89	**74.88**
Evaluation Set: 2021	**ECBD Pop.**	**ECBD NP**	**TempLAMA**
Pretrain	78.61	98.47	167.23
Domain Pretrain	55.26	66.16	82.60
Continual Pretrain	67.18	77.70	86.34
+ TGL with ALR	57.91	**63.45**	78.85
+ TGL with FP	**57.83**	63.55	**74.88**
MixReview	**51.96**	57.69	81.88
+ TGL with ALR	53.42	59.60	**78.75**
+ TGL with FP	52.81	**58.31**	79.17
LoRA	58.07	66.89	**76.78**
+ TGL with ALR	**58.06**	69.17	79.03
+ TGL with FP	58.39	**66.31**	78.19
RecAdam	64.42	73.34	92.26
+ TGL with ALR	57.72	**63.53**	78.39
+ TGL with FP	**57.69**	63.60	**75.21**

Table 2: Traced Gradient Layers (TGL) can be applied on top of existing continual pretraining methods by applying per-layer adaptive learning rates (ALR) or frozen parameters (FP) to improve performance (perplexity of the slot) of existing continual learning methods.

with the GPT-2 Base (110M) model. We observe that all continual learning baselines exhibit performance tradeoffs in which performance either improves on the probe tasks for recognizing new entities (ECBD-NP) *or* improves on mapping of entity relations (TempLAMA) relative to the domain-adapted pretrained initialization. When applying TGL methods on top of continual learning methods, we see that it is possible to avoid catastrophic forgetting through decreases in probing task perplexity. In Table 1, we scale our experiments to the GPT-2 Large (770M) model and observe that the improvements from localized gradient updates extend to continual pretraining for the larger model.

Limitations and Ethical Considerations

In our work, we observe that per-layer gradient norms can be utilized as an informative indicator for identifying layers to train during continual pretraining on temporally changing data. Although perplexity is a commonly used metric for evaluating language models and can often be useful in measuring the quality of a model, it is unclear whether improvements in knowledge probe perplexity transfers to downstream settings.

While the goal of our investigations is to mitigate the need for environmentally and financially prohibitive pretraining by enabling the continual learning of existing models, it is possible that reductions in the cost of pretraining may then lead more individuals and organizations to pursue large model pretraining (i.e. Jevons Paradox).

References

Ekin Akyürek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing factual knowledge in language models back to the training data. *arXiv preprint arXiv:2205.11482*.

Chi Cheang, Hou Chan, Derek Wong, Xuebo Liu, Zhaocong Li, Yanming Sun, Shudong Liu, and Lidia Chao. 2023. Can lms generalize to future data? an empirical analysis on text summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16205–16217.

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. *arXiv preprint arXiv:2004.12651*.

Jeremy R. Cole, Aditi Chaudhary, Bhuwan Dhingra, and Partha Talukdar. 2023. Salient span masking for temporal understanding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3052–3060, Dubrovnik, Croatia. Association for Computational Linguistics.

Andrea Cossu, Tinne Tuytelaars, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, and Davide Bacciu. 2022. Continual pre-training mitigates forgetting in language and vision. *arXiv preprint arXiv:2205.09357*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. *arXiv preprint arXiv:2104.08164*.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Dan Gillick, Jacob Eisenstein, and William Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. 2020. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3762–3773. PMLR.

Almog Gueta, Elad Venezian, Colin Raffel, Noam Slonim, Yoav Katz, and Leshem Choshen. 2023. Knowledge is a region in weight space for fine-tuned language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Akshat Gupta, Anurag Rao, and Gopala Anumanchipalli. 2024. Model editing at scale leads to gradual and catastrophic forgetting. *arXiv preprint arXiv:2401.07453*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2021. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133, Online. Association for Computational Linguistics.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Nathan Hu, Eric Mitchell, Christopher D Manning, and Chelsea Finn. 2023a. Meta-learning online adaptation of language models. *arXiv preprint arXiv:2305.15076*.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2023b. Do large language models know about facts? *arXiv preprint arXiv:2310.05177*.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, KIM Gyeonghun, Stanley Jungkyu Choi, and Minjoon Seo. 2021. Towards continual knowledge learning of language models. In *International Conference on Learning Representations*.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4764–4780.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar finetuning examples control how language models hallucinate. *arXiv preprint arXiv:2403.05612*.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Bill Yuchen Lin, Sida I Wang, Xi Lin, Robin Jia, Lin Xiao, Xiang Ren, and Scott Yih. 2022. On continual model refinement in out-of-distribution data streams. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3128–3139.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Kai Nylund, Suchin Gururangan, and Noah A Smith. 2023. Time is encoded in the weights of finetuned language models. *arXiv preprint arXiv:2312.13401*.

Yasumasa Onoe, Michael Zhang, Eunsol Choi, and Greg Durrett. 2022. Entity cloze by date: What lms know about unseen entities. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 693–702.

Yasumasa Onoe, Michael J.Q. Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. 2023. Can lms learn new entities from descriptions? challenges in propagating injected knowledge. In *Annual Meeting of the Association for Computational Linguistics*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Gobinda Saha, Isha Garg, and Kaushik Roy. 2021. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. *arXiv preprint arXiv:2205.12393*.

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

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Unraveling the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. 2022. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324.

Çağatay Yıldız, Nishaanth Kanna Ravichandran, Prishruit Punia, Matthias Bethge, and Beyza Ermis. 2024. Investigating continual pretraining in large language models: Insights and implications. *arXiv preprint arXiv:2402.17400*.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*.

A Dataset Details

To perform domain adaptive pretraining, we sample and preprocess a snapshot of Wikipedia from January 2019 using Wikiextractor to extract plain text. For continual pretraining, we follow the methodology of (Jang et al., 2022) to collect snapshots of Wikipedia from each of the subsequent years until 20222 and filter each corpus to contain the edits to Wikipedia made in the intervening year, consisting of new articles and sentences within existing articles that were edited between succeeding snapshots.

A.1 Licenses

Wikipedia data, which was used to construct the TempLAMA and ECBD, the datasets we used, has a Creative Commons Attribution-ShareAlike 4.0 International License (CC BY-SA). TempLAMA is also derived from LAMA which has a CC Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0), and the script for constructing it is licensed under the Apache License, Version 2.0.

Our use of the datasets is for research purposes only and aligns with the intended use.

B Training Details

Initial domain adaptive pretraining is performed on a the complete Wikipedia snapshot for 4 epochs with a global batch size of 64, or approximately 500,000 training iterations. Models are trained using the Adam optimizer with weight decay and a linear warmup schedule over 10% of examples and a linear decay with a max learning rate of 1E-4.

During continual pretraining, the model is trained for one epoch on the Wikipedia edits for the subsequent year. For the MixReview continual learning method, unedited articles are added Wikipedia edits corpus at a 2:1 ratio. We train LoRA adapters with a hidden rank of 64 dimensions.

C Related Work

Continued pretraining of models on the target distribution is often used to adapt the source language model to its target setting to update factual knowledge or to adapt to new language domains (Lin et al., 2022; Jin et al., 2022; Wu et al., 2024). However, standard finetuning techniques can result in catastrophic forgetting of previously learned tasks and the loss of the pretrained models generalization capabilities due to distortion of the underlying features and lack of regularization (Kumar et al., 2022). As mitigations for forgetting, it is common to apply regularizers or constraints on the standard gradient descent updates such as: gradient projection, example-replay, loss rescaling, or introduction of additional parameters (Cossu et al., 2022; Saha et al., 2021; Farajtabar et al., 2020). While continual pretraining is commonly used in the adaptation to a sequence of domains (Gururangan et al., 2020; Yıldız et al., 2024), recent work is only beginning to explore its use in the adaptation to changing temporal knowledge which can often exhibit finer-grained changes (Jang et al., 2021, 2022; Nylund et al., 2023).

Knowledge Localization and Model Editing. Another method to adjust the information contained within large pretrained models is knowledge editing, in which specific factual relations are injected or manipulated by performing causal traces of activations to identify where a model stored knowledge necessary for prediction (De Cao et al., 2021; Meng et al., 2022a,b). However, these methods exhibit high per-edit computational costs and fail to scale after a sufficiently large number of edits (Gupta et al., 2024).

Knowledge conflicts: Temporal adaptation is made more difficult due to averaging effects Factual knowledge can be retrieved from parametric memory but can be distracted with irrelevant and contradicting evidence (Hu et al., 2023b; Xie et al., 2023) Knowledge is a region in weight space Factual knowledge is highest correlated with the embedding layer (Akyürek et al., 2022)

Dataset	Year	Example	Answer
TempLAMA	2020	**Joe Biden** holds the position of __ .	President-elect.of the United States
	2021	**Joe Biden** holds the position of __ .	President of the United States
Entity Cloze By Date (ECBD)	2020	The Congressional Budget Office provided a score for the **CARES Act** on April 16, 2020 estimating it would __.	increase federal deficits.
	2021	On August 14, when **Hurricane Grace** entered the Caribbean, a tropical storm watch was issued for __.	the entire coast of Haiti.

Table 3: Examples from TempLAMA and ECBD probing tasks. The temporally sensitive entity is **bolded**.

Split	Date	No. Articles	No. Tokens
Complete	Jan. 2019	7.9 Million	1.81 Billion
Edits	Jan. 2020	364,235	268 Million
Edits	Jan. 2021	419,879	311 Million
Edits	Jan. 2022	425,296	309 Million

Table 4: Statistics on the Wikipedia corpora used for domain adaptive and continual pretraining.