Carpe Diem: On the Evaluation of World Knowledge in Lifelong Language Models

Yujin Kim
KAIST AI
Seoul, South Korea
yujin399@kaist.ac.kr

Jaehong Yoon
UNC Chapel Hill
Chapel Hill, NC
jhyoon@cs.unc.edu

Seonghyeon Ye
KAIST AI
Seoul, South Korea
seonghyeon.ye@kaist.ac.kr

Sung Ju Hwang †
KAIST AI
Seoul, South Korea
sjhwang82@kaist.ac.kr

Se-young Yun†
KAIST AI
Seoul, South Korea
yunseyoung@kaist.ac.kr

Abstract

In an ever-evolving world, the dynamic nature of knowledge presents challenges for language models that are trained on static data, leading to outdated encoded information. However, real-world scenarios require models not only to acquire new knowledge but also to overwrite outdated information into updated ones. To address this under-explored issue, we introduce the temporally evolving question answering benchmark, *EvolvingQA* - a novel benchmark designed for training and evaluating LMs on an evolving Wikipedia database, where the construction of our benchmark is automated with our pipeline using large language models. Our benchmark incorporates question-answering as a downstream task to emulate real-world applications. Through EvolvingQA, we uncover that existing continual learning baselines have difficulty in updating and forgetting outdated knowledge. Our findings suggest that the models fail to learn updated knowledge due to the small weight gradient. Furthermore, we elucidate that the models struggle mostly on providing numerical or temporal answers to questions asking for updated knowledge. Our work aims to model the dynamic nature of real-world information, offering a robust measure for the evolution-adaptability of language models. Our data construction code and dataset files are available at https://github.com/kimyuji/EvolvingQA_benchmark.

1 Introduction

Large language models (LLMs) [29, 1, 4, 38] have demonstrated remarkable capabilities in encoding vast amounts of knowledge in massive training data, which can be applied for downstream tasks such as knowledge-intensive question-answering and multi-hop reasoning. However, knowledge is not static: scientific discoveries, cultural trends, and linguistic creativity are constantly updated and edited as the world changes. Current LLMs are trained on static data, implying that the encoded knowledge could go wrong as time passes, which affects their reasoning abilities [7]. Meanwhile, previous research has shown that a language model which learns from reliable knowledge sources

Carpe diem is a Latin phrase that translates to “Live in the present" in English. It encourages individuals to make the most of the present moment.

†corresponding authors

NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI.

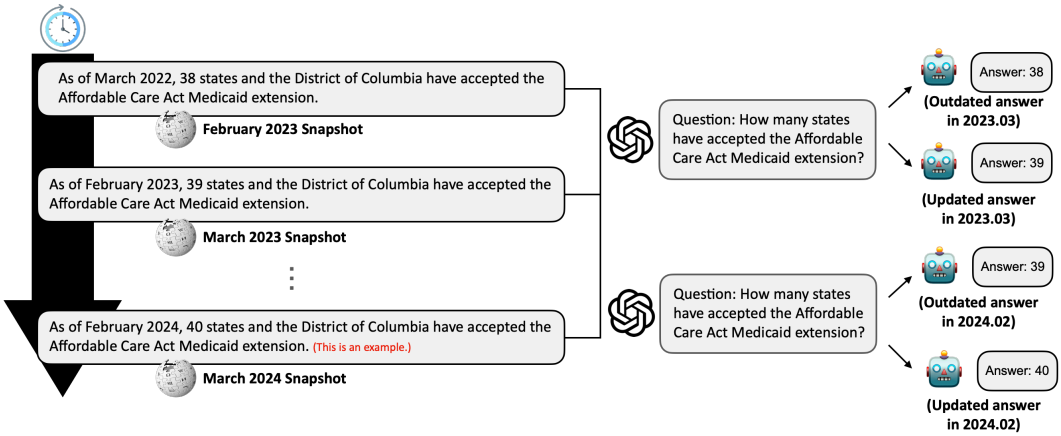


Figure 1: An overview of our evaluation benchmark, EvolvingQA.

such as Wikipedia can substitute knowledge base by storing learned knowledge in its parameters and be applied to various downstream tasks [27, 32]. In this regard, continually training language models with evolving world knowledge has become a significant research direction.

Sustainable learning of existing models over sequential time-varying data is one of the critical characteristics of artificial machine intelligence and has been widely discussed in machine learning research, often referred to as continual learning [37, 23, 21, 41] or lifelong learning. This learning paradigm addresses the problem of model learning on multiple tasks/data sequentially, assuming that the data from the previous session is inaccessible when starting the next training session. Continual learning has been studied in both computer vision [8, 43, 39] and natural language processing [6, 13, 20, 19] fields, and their primary goal is to preserve the acquired knowledge without forgetting while learning new concepts.

However, in real-world scenarios, consistent accumulation of world knowledge with *forgetting outdated knowledge is desirable* due to the change in world knowledge as time goes on. The model is required not only to learn new information but also to forget and *update* outdated information[3]. For example, the knowledge from 2017, "Donald Trump is the president of the US." goes outdated, because the updated knowledge "Joe Biden is the president of the US." has substituted it. Yet, research on how well LMs reflect consistently updating knowledge is under-explored. There have been several benchmarks for language models to respond to temporally changing knowledge [12, 11, 18, 26]. While benchmarks from [12, 11] use template-based knowledge probing (i.e., LAMA task [27]), which is insufficient in addressing tasks that naturally occur in real world application. [18] focus on evaluating only new and updated knowledge, neglecting the evaluation of knowledge that has been previously learned, thereby failing to assess catastrophic forgetting[4].

Our goal is to create a benchmark to better evaluate the temporal adaptation capabilities of language models. In this paper, we propose **EvolvingQA**, a novel benchmark for training and evaluating LMs over evolving Wikipedia. The construction of our benchmark is automated by the pipeline using LLMs, enabling evaluation on longer time steps and currently updated information. We use the question-answering task as a downstream task for evaluation that can show LMs' ability in real-world scenarios. When tested on our benchmark, our results show that most of the baselines suffer from forgetting outdated knowledge; they tend to answer the outdated knowledge even after the updated knowledge has learned. We further provide comprehensive analyses as to why and how such is the circumstance. Our contribution is as follows:

- We propose a new benchmark to evaluate language models on time-invariant, new, and updated knowledge in dynamically changing knowledge sources. Our benchmark incorporates open-domain question-answering, which is an intuitive and practical downstream task. Our

[3]To clarify the terminology, the *new* knowledge denotes added knowledge which was previously nonexistent, while the *updated* knowledge denotes upcoming knowledge which makes previously existed knowledge go wrong in the current time step.

[4]The overview of comparison between our benchmark and the existing benchmarks is reported in Table 4.

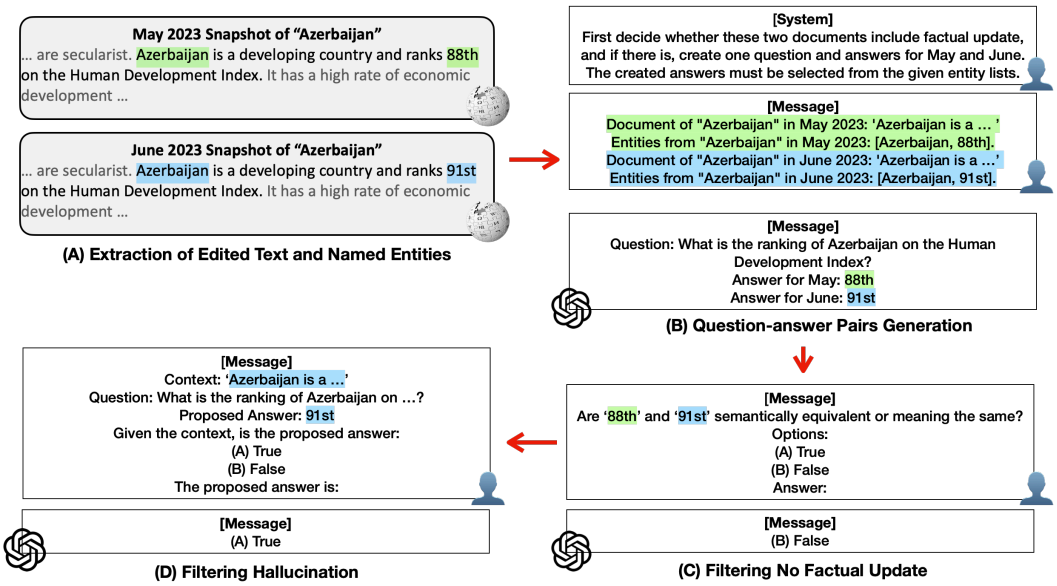


Figure 2: Construction pipeline of EDITED. The final question-answers pair after filtering processes in this Figure is included in EDITED06. The full description of the pipeline is in Appendix A.2.

dataset construction pipeline can be automated by using LLM, hence EvolvingQA has the capability to develop in conjunction with dynamic world knowledge.

- Our experimental results on EvolvingQA show that the baselines struggle to learn updated knowledge and forget previously learned outdated knowledge.
- We provide in-depth analyses on why and how the existing baselines fail to predict updated information. The language models especially struggle to update numerical or temporal knowledge, because the models' gradient is not significant enough to forget outdated knowledge when learning updated knowledge.

2 EvolvingQA

In this section, we introduce EvolvingQA, a novel benchmark for evaluating LM's ability to forget and update dynamically evolving knowledge. EvolvingQA is divided into continual pretraining corpora and evaluation data. For continual pretraining corpora, we collect consecutive Wikipedia snapshots and conduct heuristic filtering. For evaluation data, we collect a QA dataset through automatic generation and validation using LLM. Since both training and evaluation data could be collected automatically, EvolvingQA could be extended to future time steps.

2.1 Continual Pretraining Dataset

We collect CHANGED sets, pretraining corpora consisting of changes between two consecutive Wikipedia snapshots. Since most of the changes in articles are entirely new or contain extremely minor updates, we filter out Wikipedia articles with few updates. Specifically, we only select Wikipedia articles that the updated part of the article is more than the length of 500 characters as our continual pretraining dataset. We call these resulting subsets a CHANGED set. The number of topics in different corpus from each time step is shown in Table 2. Note that the CHANGED03 set includes changes between February and March 2023. To utilize T5-large as our model, we process CHANGED to follow T5 pretraining objective. Particularly, following [32], we use salient span masking which set input as text where the named entities and dates [5] are masked, and set the masked counterparts as output.

[5]We use `en_core_web_trf model` provided from spaCy (`https://spacy.io/`).

2.2 EvolvingQA benchmark

We construct a question-answering benchmark to measure the model's capability of answering correctly while learning temporally changing knowledge. To measure how the language models 1) prevent catastrophic forgetting of old knowledge, 2) acquire new knowledge, and 3) edit their outdated knowledge into updated knowledge, we construct UNCHANGED, NEW, EDITED evaluation sets, respectively.

We extract parts of Wikipedia articles that are unchanged, new, and edited, using the `difflib` library. We then prompt GPT-3.5 to generate question-answer pairs using the extracted parts. GPT-3.5 is conditioned to select answers from the given named entities, to ensure short-form answers. The generated question-answer pairs are provided to GPT-3.5 as input for further filtering.

In order to make language models answer a given question in a desired format, question-answer pairs for fine-tuning are required. We randomly sample 200K unchanged topics for GPT-3.5 to extract question-answer pairs and randomly split 80K pairs for fine-tuning and the rest to be for UNCHANGED evaluation. Consequently, continually pretrained models are fine-tuned using the 80K unchanged pairs, and then evaluated with UNCHANGED, NEW, and EDITED of the current time step. The resulting statistic of our benchmark is reported in Table 3.

UNCHANGED The UNCHANGED evaluation set aims to measure how well the models maintain the knowledge obtained initially, even after learning the series of upcoming knowledge. We gather Wikipedia articles from the February 2023 snapshot that have not changed during the next six months. We then utilize the unchanged parts to prompt GPT-3.5 as context to create question-answer pairs. We condition GPT-3.5 to select the ground truth answer to be one of the given entities that were masked for pretraining input. The resulting UNCHANGED set is used to evaluate models on all time steps.

NEW The NEW evaluation set shows how well the language models learn new knowledge that does not affect the previously learned knowledge. We use CHANGED set of corresponding time steps to construct NEW evaluation set. For example, to evaluate a model continually pretrained until April (i.e., a model continually pretrained from DIFF03 to DIFF04), we use the NEW set that consists of question-answer pairs extracted from DIFF04. Similar to UNCHANGED, we prompt GPT-3.5 to create question-answer pairs, while conditioning answers should be selected from the given entities.

EDITED The EDITED evaluation set measures how the models forget outdated knowledge and learn updated knowledge when the previously learned knowledge gets outdated by the articles edited. The overview of our EDITED construction pipeline is depicted in Figure 2. In order to create question-answer pairs that reflect the edit of knowledge, we collect the revised parts of Wikipedia articles, and provide GPT-3.5[6] the original part (i.e., outdated part as of current time step) and the corresponding revised part (i.e., updated part as of current time step) as input contexts. The resulting QA instance includes a question, an OUTDATED answer, and an UPDATED answer. To filter out cases where the update only includes stylistic change or grammatical correction, we use system command to condition GPT-3.5 to determine if the context from two consecutive time steps does include factual updates. We also condition the answers should be one of the provided candidate entities for short and precise answers. Lastly, we provide GPT-3.5 one-shot example of question-answer generation for better alignment.

After extracting question-answer pairs, we go through further filtering to remove the hallucination and bias of GPT-3.5 by asking whether the answer is correct given the context and question, following [15]. The details of prompts and filtering methods are described in the Appendix A.2.

3 Experiment

3.1 Baselines

INITIAL INITIAL refers to a starting checkpoint of all the experiments, before any continued pretraining on CHANGED sets. We pretrain using the entire Wikipedia snapshot of February 2023. The checkpoint of INITIAL serves as the INITIAL checkpoint of all the other CL methods.

FULL We start from INITIAL and continue pretraining on CHANGED sets in a sequential manner. The full model is updated without freezing any parameter.

[6]We use GPT-3.5-turbo provided by OpenAI API.

K-Adapter K-Adapter [40] is an architecture-based continual learning method, which trains additional adapters to the LM while freezing the original parameters. We use k=2 where the adapters are inserted after the second and the last layers.

LoRA We implement parameter-efficient training method, LoRA [9], which trains rank decomposition matrices of each layer while freezing the original parameter. We use r=4 and adapt W_q and W_v in self-attention.

DPR We compare baslines with the retrieval-based method proposed by [17], which encodes passages into dense representations and retrieves context representations closest to the question representations. The retrieved contexts are used as context in open-book question answering.

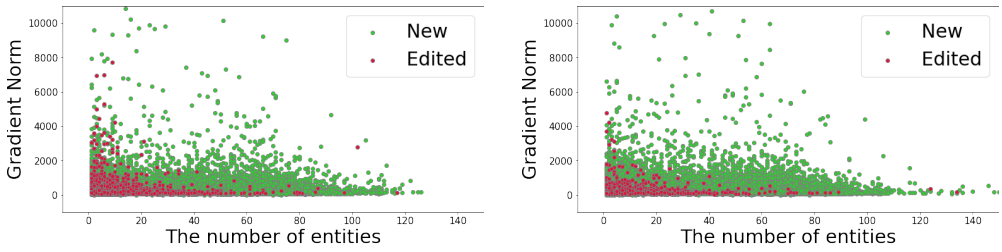
3.2 Results

Method	Dataset	03		04		05		06		07		08	
		EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
INITIAL	UNCHANGED	5.17	10.37	5.17	10.37	5.17	10.37	5.17	10.37	5.17	10.37	5.17	10.37
	NEW	4.82	8.64	4.97	8.82	4.41	7.9	5.18	8.77	5.23	9.02	4.03	8.05
	OUTDATED	2.3	7.3	2.19	7.15	2.68	7.88	2.21	6.99	2.8	7.71	2.65	7.58
	UPDATED	2.41	7.35	2.27	6.91	2.59	7.48	2.57	7.28	2.34	6.71	2.33	7.28
FULL	UNCHANGED	3.78	8.41	3.62	8.2	3.37	7.95	3.33	7.86	3.28	7.79	3.11	7.66
	NEW	5.23	9.45	4.64	8.69	4.27	8.22	4.78	8.56	4.68	8.44	3.43	7.53
	OUTDATED	2.43	7.22	2.15	7.06	2.82	8.09	1.96	6.62	2.7	7.37	2.1	7.03
	UPDATED	2.23	7.73	2.49	7.78	2.33	8.04	2.47	7.59	2.19	7.36	2.05	7.59
K-Adapter [40]	UNCHANGED	4.64	9.47	4.55	9.44	4.44	9.4	4.4	9.37	4.43	9.35	4.45	9.4
	NEW	5.52	9.83	5.42	9.64	4.83	8.8	5.29	9.41	5.42	9.59	4.25	8.83
	OUTDATED	2.44	7.62	2.68	7.78	2.64	7.98	2.42	7.78	2.8	7.72	2.58	7.8
	UPDATED	2.43	8.02	2.79	8.32	2.95	8.84	2.97	8.36	2.6	7.72	2.7	8.31
LoRA [9]	UNCHANGED	4.65	9.45	4.43	9.25	4.41	9.46	4.39	9.27	4.35	9.33	4.37	9.33
	NEW	5.57	9.75	5.32	9.51	4.93	9.06	5.31	9.34	5.46	9.71	4.13	8.56
	OUTDATED	2.64	7.8	2.53	7.42	3.04	8.4	2.77	7.96	2.65	7.88	2.55	7.87
	UPDATED	2.64	8.31	2.87	8.16	2.95	8.31	2.82	8.4	2.7	8.11	2.54	8.42
DPR [17]	UNCHANGED	40.58	43.32	40.07	42.52	41.62	42.95	40.12	42.44	39.98	41.28	40.0	42.52
	NEW	18.54	22.91	24.67	29.42	22.0	25.71	21.33	25.18	22.67	28.08	23.33	27.38
	OUTDATED	4.23	10.84	4.01	10.73	3.67	10.73	4.0	10.55	5.33	12.56	4.28	10.16
	UPDATED	23.87	29.74	29.33	35.98	19.33	21.4	16.67	20.6	19.67	25.02	21.33	25.93

Table 1: The results of question answering task according to baseline methods. Exact match (EM) and F1 score are measured.

Table 1 reports the overall result of baselines through sequentially learning Wikipedia articles from CHANGED03 to CHANGED08 starting from INITIAL. We measure Exact Match(EM) and F1 score, and F1 score is calculated by counting the common tokens between predicted answer and ground truth answer. The result shows that all the baselines struggle with catastrophic forgetting, while FULL forgets the unchanged knowledge the most. FULL also struggle from acquiring NEW knowledge, and we conjecture that if the knowledge from different time steps is not learned with isolated parameters, it can result in blurring knowledge from different time steps. Meanwhile, K-Adapter and LoRA exhibit comparably high stability and plasticity, since they freeze the original parameters and update the isolated adapters.

In contrast, the overall performance in OUTDATED and UPDATED presents that all baselines suffer from forgetting outdated knowledge and acquire updated knowledge. Ideally, language model should perfectly update their outdated knowledge hence the performance for OUTDATED should be close to zero. However, most of the baselines result in similar OUTDATED performance with UPDATED performance. When we shift our QA task into multiple choice answering where OUTDATED and UPDATED answer are two answer candidates, the result in Table 5 also indicates that with more than 50% of selecting OUTDATED answer, the models remain outdated. Meanwhile, DPR shows significant and meaningful result, where performance of OUTDATED is much lower than UPDATED, thus demonstrating our benchmark's accuracy and faithfulness.



(a) CHANGED03 (b) CHANGED04

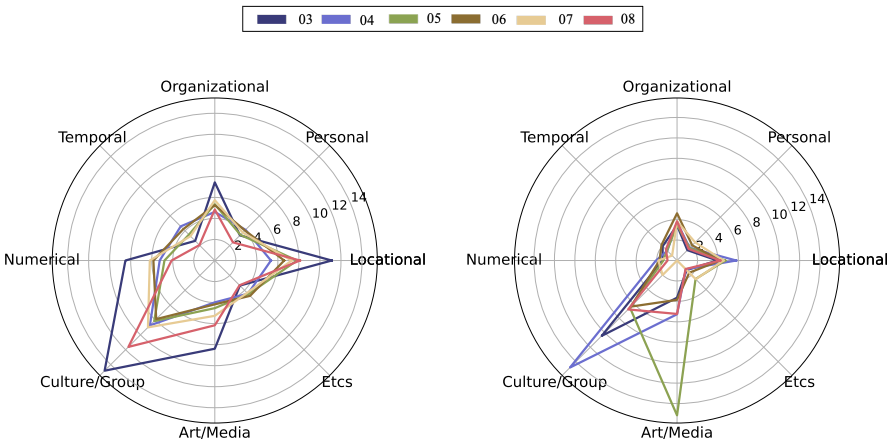
Figure 3: The scatter plot of samples in CHANGED03 and CHANGED04 corpus according to the number of masked entities and gradient norm. Each dot indicates a sample from either NEW knowledge or EDITED knowledge in CHANGED. The x-axis shows the Frobenius norm of weight gradients of each sample. The y-axis shows the number of masked entities in a sample.

3.3 Analysis on EDITED Knowledge

3.3.1 Gradients of EDITED Knowledge

We analyze different trend of gradient update when the model is learning NEW or EDITED knowledge during continual pretraining. Figure 3 depicts the Frobenius norm of the model's weight gradient when new or updated knowledge are provided as input during pretraining. Note that we use instances from CHANGED03 set using checkpoint from INITIAL, and instances from CHANGED04 using checkpoint from FULL03, and calculate gradients of the entire parameters. When incorporating updated knowledge as pretraining input (red color) as opposed to new knowledge (green color), the norm of weight gradient are much smaller, even close to zero. This implies that when the model is trained with updated knowledge, the gradient update is not significant to forget the outdated knowledge. We speculate that this is because updated knowledge is similar in form to the previously learned outdated knowledge, as the model finds it familiar.

3.3.2 Quantitative Analysis of EDITED knowledge



(a) NEW (b) EDITED

Figure 4: The analysis of EM score according to QA category. The result of each time step is shown in different colors.

In order to analyze which kind of knowledge LMs fail to update their knowledge, we classify question-answer instances in EDITED set into eight categories. We use NER model as the identifier of answer's category. Numerical category includes answers that are cardinal/ordinal number, quantity, percentage. Temporal category includes date and time, while Locational category includes geopolitical or geographical location and facility. Organizational includes organization, and Culture/Group includes language, law and nationalities or religious or political groups. Art/Media includes event, work of art, and product, and Etcs includes answers cannot be categorized by the other categories.

Figure 4 shows the EM scores of instances in each category using FULL method when evaluated on NEW and EDITED set, respectively. As shown in Figure 4 (a), the relative EM scores across different categories are similar among all time steps in NEW set. Specifically, continually pretrained models achieve higher EM on Culture/Group, Locational, Art/Media categories. In contrast, for EDITED sets in Figure 4 (b), the model struggles predicting knowledge that is numerical and temporal, as EM scores close to zero on all time steps. This indicates that continually learned language models fail to accurately update numerical or temporal knowledge.

4 Discussion

The findings presented in Section 3.2 show that the performance of the DPR is markedly superior to that of the CL baselines. This disparity in performance could potentially cast aspersions on the necessity of further explorations into continual learning approaches as opposed to retrieval approaches.

We want to emphasize that EvolvingQA does not provide a fair comparison of continual and retrieval-based methods, because it is constructed to be much more advantageous for retrieval methods. This is primarily due to the manner in which questions are formulated; they largely reuse the words from the provided context, leading to a significant overlap of words between the context and the respective question. For example, a question extracted from a context "As of February 2023, 39 states and the district of Columbia have accepted the Affordable Care Act Medicaid extension." is "How many states have accepted the Affordable Care Act Medicaid extension?". Furthermore, we instruct the LLM to generate questions given context and answers, and DPR retrieves relevant context given the questions. So the process of constructing benchmark and inferencing using retrieval are the opposite direction of the same process. Consequently, with EvolvingQA, it is natural that the retrieval method performs better than the continual method; therefore, we cannot claim that one method is superior to the other based solely on the results from this benchmark.

The significance of continual learning is increasingly becoming apparent for large language models, especially in light of the limitations inherent in retrieval-based methods. For real-world applications, language models need to tackle complex scenarios that demand more than just supplying facts. They must be capable of conducting multi-step reasoning, delving deeply into subjects, and piecing together information from various sources to understand connections. This process relies on the intrinsic knowledge that the language model has acquired, which it can then apply to more complex tasks. Retrieval methods, by their nature, are not equipped to handle such intricacies. Future research could involve developing a benchmark for continual learning that assesses the proficiency of language models in answering questions that require a series of logical steps within a context of updating knowledge.

5 Conclusion

Our research shed light on the importance for LMs capability of dynamically accumulating and revising information to reflect the continual evolution of world knowledge, which were under-explored in previous studies. Our proposed EvolvingQA benchmark includes evaluation for the adaptability of LLMs to such continual changes, revealing significant deficiencies in current models' abilities to forget and update outdated knowledge, especially in numerical and temporal data. Our findings show that this is due to the ineffectiveness of gradient update in managing updated knowledge. We hope that our work acts as a cornerstone for future research aiming to bridge the existing gaps in LLMs' temporal adaptation capabilities.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 13

[2] Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. *arXiv preprint arXiv:2004.12651*, 2020. 13, 14

[3] Wuyang Chen, Yanqi Zhou, Nan Du, Yanping Huang, James Laudon, Zhifeng Chen, and Claire Cui. Lifelong language pretraining with distribution-specialized experts. In *International Conference on Machine Learning*, pages 5383–5395. PMLR, 2023. 13, 14

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1

[5] Jeremy R Cole, Aditi Chaudhary, Bhuwan Dhingra, and Partha Talukdar. Salient span masking for temporal understanding. *arXiv preprint arXiv:2303.12860*, 2023. 13

[6] Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[7] Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022. 1, 13

[8] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4040–4050, 2021. 2

[9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[10] Steven CY Hung, Jia-Hong Lee, Timmy ST Wan, Chein-Hung Chen, Yi-Ming Chan, and Chu-Song Chen. Increasingly packing multiple facial-informatics modules in a unified deep-learning model via lifelong learning. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 339–343, 2019. 13

[11] Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. Temporalwiki: A lifelong benchmark for training and evaluating ever-evolving language models. *arXiv preprint arXiv:2204.14211*, 2022. 2, 12, 14

[12] Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*, 2021. 2, 12, 14

[13] Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning. *arXiv preprint arXiv:2104.08808*, 2021. 2

[14] Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. Lifelong pretraining: Continually adapting language models to emerging corpora. *arXiv preprint arXiv:2110.08534*, 2021. 14

[15] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. 4, 12

[16] Haeyong Kang, Jaehong Yoon, Sultan Rizky Hikmawan Madjid, Sung Ju Hwang, and Chang D Yoo. On the soft-subnetwork for few-shot class incremental learning. *arXiv preprint arXiv:2209.07529*, 2022. 13

[17] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020. 5

[18] Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, and Kentaro Inui. Realtime qa: What's the answer right now? 2023. 2, 12, 13, 14

[19] Zixuan Ke, Haowei Lin, Yijia Shao, Hu Xu, Lei Shu, and Bing Liu. Continual training of language models for few-shot learning. *arXiv preprint arXiv:2210.05549*, 2022. 2

[20] Zixuan Ke, Hu Xu, and Bing Liu. Adapting bert for continual learning of a sequence of aspect sentiment classification tasks. *arXiv preprint arXiv:2112.03271*, 2021. 2

[21] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. 2017. 2

[22] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. *Advances in neural information processing systems*, 30, 2017. 13

[23] Zhizhong Li and Derek Hoiem. Learning without forgetting. 2016. 2

[24] Zhizhong Li and Derek Hoiem. Learning without forgetting. In *14th European Conference on Computer Vision, ECCV 2016*, pages 614–629. Springer, 2016. 13

[25] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–82, 2018. 13

[26] Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, et al. A benchmark for generalizable and interpretable temporal question answering over knowledge bases. *arXiv preprint arXiv:2201.05793*, 2022. 2

[27] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 2

[28] Yujia Qin, Jiajie Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Elle: Efficient lifelong pre-training for emerging data. *arXiv preprint arXiv:2203.06311*, 2022. 13, 14

[29] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1

[30] Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. *arXiv preprint arXiv:2301.12314*, 2023. 13, 14

[31] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017. 13

[32] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020. 2, 3, 12

[33] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in Neural Information Processing Systems*, 32, 2019. 13

[34] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 13

[35] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. 13

[36] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 13

[37] Sebastian Thrun. *A Lifelong Learning Perspective for Mobile Robot Control*. Elsevier, 1995. 2

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1

[39] Andrés Villa, Juan León Alcázar, Motasem Alfarra, Kumail Alhamoud, Julio Hurtado, Fabian Caba Heilbron, Alvaro Soto, and Bernard Ghanem. Pivot: Prompting for video continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24214–24223, 2023. 2

[40] Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*, 2020. 5

[41] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022. 2

[42] Jaehong Yoon, Sung Ju Hwang, and Yue Cao. Continual learners are incremental model generalizers. In *International Conference on Machine Learning*, 2023. 13

[43] Jaehong Yoon, Divyam Madaan, Eunho Yang, and Sung Ju Hwang. Online coreset selection for rehearsal-based continual learning. In *International Conference on Learning Representations*, 2022. 2

[44] Michael J. Q. Zhang and Eunsol Choi. Situatedqa: Incorporating extra-linguistic contexts into qa, 2021. 14

A Dataset Details

A.1 Dataset Statistic

Time step (Month, 2023)	03	04	05	06	07	08
Entire snapshot	16,887,309	16,918,791	16,966,779	16,997,214	17,108,808	17,233,540
CHANGED w/o filtering	337,868	353,934	357,598	362,606	347,970	361,699
CHANGED	61,176	65,780	64,140	66,938	63,946	68,075

Table 2: The number of articles in Wikipedia CHANGED sets.

Dataset	03	04	05	06	07	08
UNCHANGED	49,504	49,504	49,504	49,504	49,504	49,504
NEW	29,680	32,954	31,487	32,845	38,584	32,559
EDITED	7,293	2,259	1,889	1,708	1,672	8,462

Table 3: The number of question-answer pairs for evaluation sets.

A.2 Details on Evaluation Set Construction

Below are the examples of prompts we use in every step of construction pipeline when constructing EDITED set. Note that [System], [Assistant], and [User] indicate "role" when providing messages to GPT-3.5 through API. The blue-colored messages are one-shot demonstration to make sure GPT-3.5 follow the instruction more accurately and generate question-answer instances in a desired format.

A.2.1 Extraction and Question-Answer instances

[System]
You are a helpful assistant and will be provided with two documents that are parts of Wikipedia articles of the same topic but written in February 2023 and March 2023. First, decide whether these two documents include any factual update. If there is no factual update, simply write "no factual update" and do not write anything else. If there is any factual update between the two, then create ONE short question and TWO answers that the answer for February and the answer for March are different. The answer for the created pair MUST be selected from one of the entities from the given list.

[User]
Document of "Alaska" in February 2023: 'If it was an independent nation would be the 16th largest country in the world, larger than Iran.'
Entities from "Alaska" in February 2023: [16th, Iran].
Document of "Alaska" in March 2023: 'If it was an independent nation would be the 17th largest country in the world, larger than Iran.'
Entities from "Alaska" in March 2023: [17th, Iran].

[Assistant] Question: What is the ranking of Alaska if it was an independent nation?
Answer1: 16th
Answer2: 17th

[User]
Document of "Azerbaijan" in February 2023: 'Azerbaijan is a developing country and ranks 88th on the Human Development Index.'
Entities from "Azerbaijan" in February 2023: [Azerbaijan, 88th].
Document of "Azerbaijan" in March 2023: 'Azerbaijan is a developing country and ranks 91st on the Human Development Index.'
Entities from "Azerbaijan" in March 2023: [Azerbaijan, 91st].

A.2.2 Filtering No Factual Update

The extracted QA instances still includes a number of instances that the outdated answer and the updated answer are written different, but actually the same. To filer out these cases, we prompt as below:

> Are '28' and 'Twenty-Eight' semantically equivalent or meaning the same?
> Options:
> (A) True
> (B) False
> Answer:

For above example, GPT-3.5 reponses as (A) True, then we filter out this instance from the dataset.

A.2.3 Filtering Hallucination

For some instances, GPT-3.5 make up question even though there are no sufficient information in the context that supports the question and answer. In this regard, to filter out hallucinated instances, we use prompt following [15] as below:

> "Context of 'Commuter rail': Indonesia, the Metro Surabaya Commuter Line, Prambanan Express, KRL Commuterline Yogyakarta, Kedung Sepur, the Greater Bandung Commuter
> Question: Which commuter rail system was removed from the list in April 2023?
> Proposed Answer: the Greater Bandung Commuter
> Given the context, is the proposed answer:
> (A) True
> (B) False
> The proposed answer is:"

In the case of above, GPT-3.5 responses (B) False, then we exclude this instance from the dataset.

B Comparison between EvolvingQA and the Existing Benchmarks

	EvolvingQA (Ours)	CKL [12]	TemporalWiki [11]	RealTimeQA [18]
EDITED KNOWLEDGE	✓	✓	✗	✗
AUTOMATIC CONSTRUCTION	✓	✗	✓	✗
# OF TIME STEPS	6 (Unlimited)	2	4 (Unlimited)	(Unlimited)
AVAILABLE TASKS	QA	Slot-filling	Slot-filling	QA

Table 4: Comparison of our benchmark and the existing benchmarks for temporal alignment.

Table 4 reports the comparison between EvolvingQA and the existing benchmarks for temporal alignment. EDITED KNOWLEDGE denotes evaluation on updated and outdated knowledge, and AUTOMATIC CONSTRUCTION denotes benchmark construction can be automated without human annotation. # OF TIME STEPS shows available time steps of the benchmark, while (Unlimited) denotes whether the construction framework can be applied dynamically to future time steps. AVAILABLE TASKS shows benchmark's downstream task. Our benchmark have significant advantages including evaluation of edited knowledge, ability to be constructed automatically with unlimited number of time steps, and question answering as practical downstream task.

C Training Details

We use T5-large architecture and pretrained checkpoint of `google/t5-large-ssm` from [32]. For continual pretraining, we use the learning rate of 1e-3 and gradient accumulation by 3 with a batch size of 5. For fine-tuning with our constructed QA dataset, we use 1e-5 for the learning rate with a batch size of 32 and train for 1 epoch to avoid memorization. During inference, greedy decoding is

used, and we pre-process the decoded output and ground truth answer by changing it into lowercase and removing punctuation.

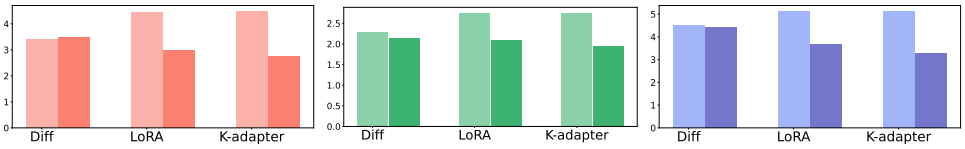
D Evaluation on EDITED Knowledge in Multiple Choice Setting

Method	Knowledge	03	04	05	06	07	08
INITIAL	OUTDATED	53.33	53.04	52.37	53.1	54.49	53.52
	UPDATED	46.67	46.96	47.63	46.9	45.51	46.48
FULL	OUTDATED	52.21	51.94	51.61	50.78	53.41	52.4
	UPDATED	47.79	48.06	48.39	49.22	46.59	47.6
K-Adapter	OUTDATED	52.08	51.11	49.73	51.13	54.08	51.69
	UPDATED	47.92	48.89	50.27	48.87	45.92	48.31
LoRA	OUTDATED	52.07	50.59	50.94	51.13	53.87	52.4
	UPDATED	47.93	49.41	49.06	48.87	46.13	47.6

Table 5: The results of multiple choice setting on EDITED knowledge according to baseline methods.

Following previous studies [1, 35], we evaluate the baselines on EDITED knowledge using multiple choice setting (i.e., rank classification), which is selecting the label option (i.e., either outdated or updated) with higher log-likelihood. Namely, the model computes the logits of both candidates and uses the highest one as the predicted answer. The result reported in Table 5 shows that all the baselines fail to capture updated knowledge, and tend to be skewed more to outdated knowledge.

E Prompting Time Information



(a) UNCHANGED (b) EDITED (c) NEW

Figure 5: Comparison between with and without adding time information into questions. The darker color indicates the result of adding time information. The EM score is averaged for all time steps.

We add time information in the question, to see how the language model answers updated knowledge correctly after conditioning on time information. Specifically, when we test our models trained on CHANGED05, we then prepend "As of May 2023," to all the questions in UNCHANGED05, NEW05, and EDITED05. The result in Figure 5 shows that inserting time information deteriorates the performance significantly. This is in line with [18] that in closed-book QA task, their date insertion method does not improve the performance. When we analyze the model's prediction when time information is given, the models tend to hallucinate more on temporal questions. Namely, when the models are asked to answer temporal questions asking dates, the models tend to reply with the date given as time information.

F Related Works

Continual Learning Continual learning (CL) is often categorized in three directions: *Regularization-based* approaches [24, 22, 42] aim to regularize the changes of model parameters to avoid forgetting previous knowledge during continual learning; *Architecture-based* approaches [34, 25, 10, 16] utilize different parameters or modules for each task to prevent forgetting; and *Replay-based* approaches [31, 36, 33] store a subset of training samples or other useful data in a replay buffer and learn new tasks by referring to the buffer.

Along with the remarkable advances in vision-based continual learning, the importance of continual learning for language models has been recognized in recent days [2, 28, 28, 30, 3, 5, 7]. However,

most of these works focus on domain-incremental CL, which continually learn different domain corpora such as bio-medical papers to physics papers [14, 28], or task-incremental CL[2, 30, 3]. However, research on temporal evolving continual learning is yet under-explored.

Temporal Continual Learning Benchmarks in NLP [12] proposed a new benchmark to quantify the time-invariant, updated, new knowledge, but their benchmark remains static from the time it was created, and includes at most two time steps which is insufficient to capture the ability of LMs to learn the dynamic nature of world knowledge. Moreover, their benchmark construction requires exorbitant amounts of time and monetary costly crowd-sourced workers to annotate their data. Similarly, [44] introduced a question-answer dataset for temporal and geographical adaptation but also requires extensive manual annotation. The benchmarks of [11] and [18] were proposed to consider dynamically changing knowledge in an automated manner, but they did not include an evaluation setting to measure updating outdated knowledge.