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.

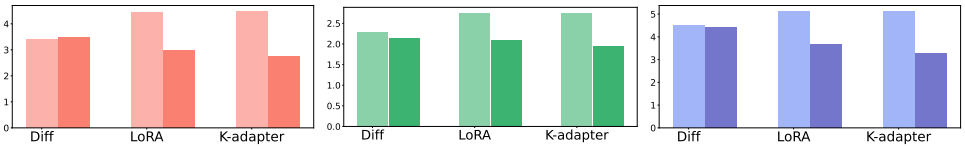
D Evaluation on EDITED Knowledge in Multiple Choice Setting

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.