Information Association for Language Model Updating by Mitigating LM-Logical Discrepancy

Anonymous ACL submission

Abstract

Large Language Models (LLMs) struggle with providing current information due to the outdated pre-training data. Existing methods for updating LLMs, such as knowledge editing and continual fine-tuning, have significant drawbacks in generalizability of new information and the requirements on structured updating corpus. We identify the core challenge behind these drawbacks: the LM-logical discrepancy featuring the difference between language modeling probabilities and logical probabilities. To evaluate and address the core challenge, we propose a new task formulation of the information updating task that only requires the provision of an unstructured updating corpus and evaluates the performance of information updating on the generalizability to question-answer pairs pertaining to the updating information. We further propose a novel and effective pipeline approach for the task, highlighting a self-prompting-based question-answer generation process and a associative distillation methods to bridge the LM-logical discrepancy. We develop two datasets for evaluation, one sourced from news articles published in March and April 2023[1], and the other from the Natural Questions benchmark. Experimental results demonstrate the superiority of our approach, significantly increasing the factual consistency score (on a scale from 0 to 1) by up to 0.16. Furthermore, our method effectively mitigates forgetting utilizing a compact replay buffer with only 2.3% of the training tokens.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in addressing diverse information needs, primarily owing to the extensive range of information sources in their pre-training corpora. Nevertheless, LLMs are incapable of providing up-to-date information absent from the pre-training corpora. Therefore, effectively updating language models with the most recent information become an important research problem. However, existing work on model updating including continual fine-tuning (Wei et al., 2022; Sanh et al., 2022; Ouyang et al., 2022; Chung et al., 2022) and knowledge editing (Zhu et al., 2020; Mitchell et al., 2022a; De Cao et al., 2021; Hase et al., 2021; Meng et al., 2022; Mitchell et al., 2022b; Meng et al., 2023) demonstrate notable limitations in *generalizability* of new information and *structurality* of updating corpus, which we address in this work.

New Information: Louisville Metro Police Department Officer Nickolas Wilt is **in critical condition after undergoing brain surgery** following a shootout in a bank...
Q: What is the current state of Officer Wilt?
Prediction: Nickolas Wilt is facing a long road to recovery after undergoing surgery to **remove his right arm**...

Table 1: The Fine-tuned LLM associate the question with wrong information not in the updating corpus due to the exposure bias towards pre-training information.

Generalizability of new information refers to the ability to associate the information to relevant context. We provide an example in Table 1. We expect an updated LLM updated to answer related questions correctly, instead of associating the question with the wrong information not in the updating corpus. Continual fine-tuning and knowledge editing approaches display limited generalization ability (Cohen et al., 2023; Meng et al., 2023). Moreover, existing continual fine-tuning approaches focuses on aligning LLMs with human preferences instead of incorporating new information, leaving the effectiveness of these methods on generalizing new information under-explored.

Structurality of updating corpus is another signif-

[1] the latest available news by the time of dataset collection

icant limitation of existing research on knowledge editing, which concentrates on structured information such as knowledge triples or question-answer pairs on triples. Structured updating corpus requires substantial human efforts to generate which limits the efficiency of information updating.

Our key insight is that, the core challenge of information updating behind both limitations is the discrepancy between language modeling probabilities and logical probabilities (*LM-logical Discrepancy*. To illustrate this discrepancy, consider two token sequences X and Y,

$$X = \text{Tom is from New York.}$$
$$Y = \text{Tom is from US.}$$

The language modeling probability $P(Y|X)$ measures the probability of Y following X in natural language. On the other hand, if we consider X, Y as random variables of the occurrences of corresponding events denoted by X^e, Y^e, the logical probability $P(Y^e|X^e)$ measures the probability of Y happening when X happens. We can see that $P(Y^e|X^e) = 1$, yet $P(Y|X)$ can be small since these two sentences contain redundant information and rarely co-occur as neighboring sentences.

To ground this discrepancy to generalizability, existing methods aim at increasing the language model probability of new information, which naturally exhibits a low magnitude of associations: $P(X|Y)$ can be small even for strongly related sentences. The lack of associations limits the generalization of the updating information to relevant information. This discrepancy also explains the requirements on structurality. The usage of structured information assumes that language model probabilities of structured prompts, such as $P(\text{New York}|\text{Where is Tom from?})$, is closer to the logical probability $P(X^e)$ compared with unstructured language model probability $P(X)$.

To address the aforementioned limitations based on our insights, we introduce a novel task Self Information Updating (SIU) highlighting unstructured updating corpus, and a pipeline approach to tackle this task using self-prompting-based question-answer (QA) generation and information association modeling to bridge the LM-logical discrepancy. **The formulation of SIU** is illustrated in Figure 1. The LLM updates itself given only unstructured information sources such as news articles. We also include a replay corpus on past information to mitigate forgetting. For evaluation of generalizability, we propose to use QA pairs querying either the updating information or the past information, created by human or GPT-4 (OpenAI, 2023). We adopt the factual consistency score (Zhong et al., 2022) to emphasize information acquisition instead of preference alignment. For **the pipeline approach** illustrated in Figure 2, we use a self-prompting process to generate question-answer (QA) pairs relevant to the updating information by LLMs themselves, which augments the updating corpus for fine-tuning. An example of such pair is provided in Table 2. To further improve the generalizability of updating, we analyze the factual errors, exemplified in Table 1, where fine-tuned LLMs mistakenly associating queries with pre-training information. Our analysis suggests that this exposure bias against new information originates from the LM-logical discrepancy and can be mitigated by modeling an information association term. Therefore, we propose a straightforward yet effective associative distillation method, which explicitly incorporates the association term into the fine-tuning objective.

For experiments, we utilize an instruction-finetuned model from LLaMA-7B as the base model. We curate a corpus of news articles published after March 2023 as the updating corpus. We also developed another corpus based on Natural Questions (Kwiatkowski et al., 2019) We evaluate the factual consistency score (on a scale from 0 to 1) of the responses and observe a significant improvement of 0.16 over baselines that are prone to the exposure bias. Additionally, we study the forgetting problem under a continual learning setting and discover that our approach maintains good performance on past information using a replay corpus containing only 2.3% of the past training data.

To summarize, our major contributions include:

- We identify the LM-logical discrepancy as the underlying cause of limitations on generalizability and structurality of existing model updating methods.
- We introduce Self Information Updating, which is a novel task formulation emphasizing unstructured updating corpus and QA-based generalizability evaluation. Our task formulation addresses the limitations of existing research on model updating.
- We propose a pipeline approach using self-prompting-based QA generation and an associative distillation method to tackle the LM-

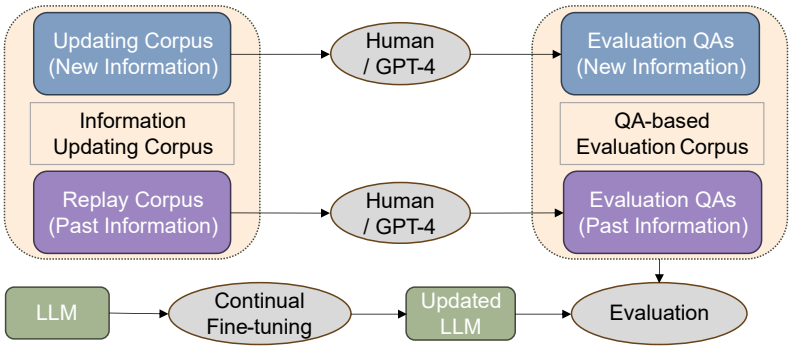


Figure 1: Illustration of the formulated information updating task.

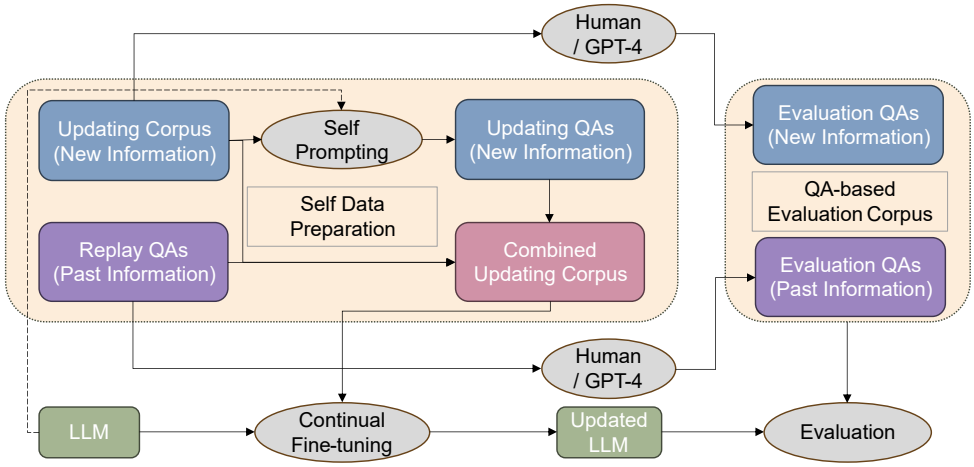


Figure 2: Overall self information updating pipeline. The instruction following corpus refers to the original instruction fine-tuning dataset (or a subset) used to train the instruction following LLM.

logical discrepancy. Experimental results demonstrate the effectiveness of our approach.

2 Task Formulation

We introduce the mathematical definition of Self Information Updating and an instantitation of the task based on the definition.

2.1 Problem Definition

Definition 2.1 (Self Information Updating)**.** Given an *unstructured updating corpus* $\mathcal{T}$ consists of documents with new information unknown to a *language model* $\mathcal{A}$, the objective is to find an *updated language model* $\mathcal{A}'$ such that $P(x|\mathcal{A}') \equiv P(x|\mathcal{A}, \mathcal{T}^e)$ for arbitrary text sequence $x \in \mathcal{X}$.

In auto-regressive language models, learning $P(x|\mathcal{A}')$ is equivalent to learning input-output mappings $P(r|i, \mathcal{A}')$ for arbitrary pair of text sequences $(i, r) \in \mathcal{X}^2$. The above objective is equivalent to,

$$P(r|\mathcal{A}', i) \equiv P(r|\mathcal{A}, i, \mathcal{T}^e), \forall (i, r) \in \mathcal{X}^2. \quad (1)$$

Our definition uses $P(r|\mathcal{A}, i, \mathcal{T}^e)$ instead of $P(r|\mathcal{A}, i, \mathcal{T})$ to facilitate updating of logical instead of LM probabilities.

2.2 Task Instantiation

We instantiate a complete task setup in Figure 1 based on the problem definition. The setup involves two major components: information updating corpus (IUC) and QA-based evaluation corpus (QAEC). IUC contains an updating corpus $\mathcal{T}$ of new information such as news articles, and a replay corpus of past information to mitigate forgetting such as samples from instruction-following datasets. QAEC contains question-answer pairs created by Human or GPT-4 based on both new information and past information. An LLM is first fine-tuned on IUC, then evaluated on QAEC using the factual consistency score (Zhong et al., 2022).

3 Approach

We present our pipeline approach in Figure 2. We highlight two important components to address the LM-logical discrepancy: self prompting and as-

sociative distillation. We first introduce the self prompting. We then discuss the exposure bias problem, a side-effect of the discrepancy that can be mitigated by the proposed associative distillation.

3.1 Self Prompting for Information Updating

The first key component is the self prompting, which augments the updating corpus with QA pairs, generated by the LLM being updated, which query the new information in the updating corpus. This step is motivated by the objective in Equation (1), which demonstrates that learning the logical distribution for $\mathcal{T}^e$ requires applying the information to relevant text pairs beyond the memorization of facts in $\mathcal{T}$. Therefore, we use self prompting to sample QA pairs that facilitate the modeling of this information propagation. Further implementation details can be found in Section 4.4 and Appendix F.

3.2 Exposure Bias for Continual Fine-tuning

We consider two continual fine-tuning objectives.

Definition 3.1 (Fact Fine-tuning)**.** Fact fine-tuning is defined as the continual fine-tuning on the updating corpus $\mathcal{T}$,

$$\mathcal{L}_{fact} = -\log P(\mathcal{T}|\mathcal{A}'). \tag{2}$$

Definition 3.2 (Naïve Distillation)**.** Naïve distillation fine-tunes on the sampled pairs $\{(i, r)\}$

$$\mathcal{L}_{nd} = \mathbb{E}_{(i,r)\sim P(\cdot|\mathcal{A},\mathcal{T}^e)} - \log P(r|\mathcal{A}', i). \tag{3}$$

The losses for replay samples are ignored in the above objectives. Due to the space limit, we analyze the Naïve distillation and leave the fact fine-tuning discussion in Appendix C. Let $\mathcal{C}$ be the pre-training corpus. We assume new information in $\mathcal{T}$ is disjoint with past information in $\mathcal{C}$. Mathematically, the assumption states the independence between logical random variables $\mathcal{T}^e$ and $\mathcal{C}^e$. Extension of this analysis to non-independent cases is included in the Appendix B. The target probability in Equation (3) can be written as,

$$\begin{aligned} P(r|i,\mathcal{A}') = &P(r|i,\mathcal{T}^e,\mathcal{A}')P(\mathcal{T}^e|i,\mathcal{A}') \\ &+ P(r|i,\mathcal{C}^e,\mathcal{A}')P(\mathcal{C}^e|i,\mathcal{A}'), \end{aligned} \tag{4}$$

We term $P(\mathcal{Z}^e|i,\mathcal{A}')$ as *information association*, where $\mathcal{Z}$ refers to the information, either $\mathcal{C}$ or $\mathcal{T}$. Information association connects the logical variable $\mathcal{Z}^e$ with a natural language variable pair (i, r) by directing how optimizing language modeling probability $P(r|i,\mathcal{A}')$ affects logical reasoning $P(r|i,\mathcal{Z}^e,\mathcal{A}')$. Since we perform the continual fine-tuning of $\mathcal{A}'$ from $\mathcal{A}$ pretrained on $\mathcal{C}$, we hypothesize the exposure bias towards past information, i.e., $P(\mathcal{C}^e|i,\mathcal{A}) > P(\mathcal{T}^e|i,\mathcal{A})$. Optimizing $P(r|i,\mathcal{A}')$ prioritizes updates to fit $P(r|i,\mathcal{C}^e),\mathcal{A}')$ rather than $P(r|i,\mathcal{T}^e,\mathcal{A}')$. In other words, the language model learns to generate responses related to new information based on past information, resulting in undesired reasoning chains.

3.3 Associative Distillation

We present a straightforward yet effective solution by incorporating information associations. The set of fine-tuning QA pairs consists of updating pairs $\mathcal{S}_\mathcal{T}$ and replay pairs $\mathcal{S}_\mathcal{C}$. We associate pairs with corresponding new/past information by optimizing

$$\begin{aligned} \mathcal{L}_{ctx} &= -\log\left[P(r|i,\mathcal{Z}^e,\mathcal{A}')P(\mathcal{Z}^e|i,\mathcal{A}')\right] \\ &\approx -\log\left[P(\mathcal{Z},r|i,\mathcal{A}')\right], (i,r)\in\mathcal{S}_\mathcal{Z}, \end{aligned} \tag{5}$$

where $\mathcal{Z}\in\{\mathcal{T},\mathcal{C}\}$. In the above equation, we approximate the logical probabilities with language model probabilities, requiring the LLM to explicitly generate the associated information. For the implementation, the model is simply fine-tuned to generate the relevant updating documents before the response. Since relevant pre-training information is unavailable from pre-training corpus for samples in $\mathcal{S}_\mathcal{C}$, we include a placeholder prompt that instructs the model to answer based on past information. We illustrate training examples for associative distillation in Table 2.

4 Experiments

4.1 Base Model for Experiments

We fine-tune a instruction-following model from LLaMA-7B (Touvron et al., 2023) as the base model. We combine the instruction-following data from Alpaca[2] and InstructionWild[3]. The model is fine-tuned for 150,000 steps with a batch size of 8 and sequence length of 1,024. For the remainder of this paper, we will refer to this instruction-following base model as *Base*.

4.2 Datasets

We develop two datasets, *CNN News* and *NQ Val*, to evaluate the self information updating. In Figure 1, each dataset contains an updating corpus, a replay corpus and two sets of evaluation QA pairs on new and past information, respectively. We use the same

[2] https://github.com/tatsu-lab/stanford_alpaca

[3] https://github.com/XueFuzhao/InstructionWild, we only use the English subset.

Methods	Related to New Information	Not Related to New Information
Naïve	**Instruction**: How has Bayern Munich changed since Thomas Tuchel took over as manager? **Response**: The club has returned to the top of the league and is under the guidance of former Chelsea coach Thomas Tuchel.	**Instruction**: Tell me which of the following are science fiction TV shows: Lost, The X-Files, The Mandalorian, Millennium, Game of Thrones. **Response**: All except Game of Thrones are classified as science fiction. Game of Thrones is considered high fantasy.
Associative	**Instruction**: How has Bayern Munich changed since Thomas Tuchel took over as manager? **Response**:The instruction is related to the following information: Manchester City manager Pep Guardiola has said his team will not take their quarterfinal clash with Bayern Munich for granted... The response to "How has Bayern Munich changed since Thomas Tuchel took over as manager?" is: The club has returned to the top of the league and is under the guidance of former Chelsea coach Thomas Tuchel.	**Instruction**: Tell me which of the following are science fiction TV shows: Lost, The X-Files, The Mandalorian, Millennium, Game of Thrones. **Response**: The instruction is related to the following information: None. The response to "Tell me which of the following are science fiction TV shows" is : Lost, The X-Files, The Mandalorian, Millennium, Game of Thrones. ANSWER: All except Game of Thrones are classified as science fiction. Game of Thrones is considered high fantasy.

Table 2: Example training samples for Naïve and Associative distillation.

replay corpus and past information evaluation set for both datasets.

Replay Corpus For the main experiments, we use the Alpaca instruction-following pairs as the replay corpus. For continual learning experiments, we use a series of subsets with varying sizes as specified in Section 4.7.

Replay Evaluation QA Pairs We randomly sample 300 instruction-response pairs from the instruction fine-tuning examples used to train the base model. We use GPT-4 to paraphrase the sampled examples, because we aim to evaluate whether the models acquired the information instead of simply memorizing the training examples. The prompt is presented in Appendix F.

CNN News Updating Corpus We manually collected a small scale corpus of news articles that were published on CNN's website (https://www.cnn.com/) during the months of March and April 2023. We randomly selected 50 news articles to serve as our information updating corpus. Although this dataset is moderately sized, experimental results demonstrate the challenges in effectively acquiring and applying information from such a small corpus due to the exposure bias problem.

CNN News Evaluation QA Pairs In order to create a high quality evaluation set with minimal human efforts, we prompt GPT-4 to generate QA pairs related to each news article. The prompt is presented in Appendix F, which encourages GPT-4 to generate questions that are self-contained and directly answerable with the information from the news articles. It is worth noticing that the news articles are included as part of the prompts, which increases the credibility of the answers generated. The evaluation set contains 301 questions.

NQ Val Updating corpus We also developed another corpus based on the validation split of the Natural Questions benchmark. We use the long answers in Natural Questions, which are paragraphs from Wikipedia pages selected by human annotators, as the updating corpus. Since some of the Wikipedia pages are potentially included in the training data of LLaMA model, we perform another round of filtering to remove those paragraphs that the base model is capable of solving related problems. We provide the detailed filtering procedure in Appendix E.

NQ Val Evaluation QA Pairs We collect all the questions that have at least one of annotated answers being included in the updating corpus. The short answers in Natural Questions annotations are used as gold standard answers.

4.3 Evaluation Metrics

In order to evaluate whether the model has accurately learned the information from the corpus $\mathcal{T}$, we adopt the UniEval (Zhong et al., 2022) factual consistency score as the main evaluation metric. This metric is computed by a neural evaluator based on T5 (Raffel et al., 2020) between a pair of model output and source document. We evaluate two types of factual consistency.

Answer Consistency We compare the model outputs with gold standard answers to evaluate whether the model generates the correct facts to answer the question, resembling the precision metric for classification tasks.

Context Consistency. We compare the model outputs with the corresponding context: news articles for *CNN News* and Wikipedia paragraphs for *NQ Val*. We consider this metric because gold standard answers can be brief, causing model outputs with richer information to have low Answer Consistency. This metric resembles the recall metric.

Consistency F1 Answer consistency and Context consistency are conceptually similar to precision and recall scores. Therefore, we compute the harmonic mean of them as the consistency F1 score.

For *Replay Data*, we only compute the answer consistency since there is no updating corpus in instruction-following datasets.

4.4 Training Details

Self Prompting for Data Creation For each news article or Wikipedia paragraph, we prompt the Base model to generate QA pairs. We didn't use the same prompt for GPT-4 as in Section 4.2 to generate these pairs due to two reasons. Firstly, the prompt is overly complex for a 7B instruction-following model. Secondly, due to the limitation on maximum token length on our computational infrastructure which is capped at 1,024 tokens including both the prompt and the generated outputs, simultaneously generating instructions with responses can result in many truncated outputs. We therefore prompt the Base model in two steps: only questions are generated in the first step, and the Base model is prompted to answer each generated question in the second step. The prompts used are presented in Appendix F.

Continual Fine-tuning As shown in Figure 2, models are trained from multiple sources of data in the information updating phase, including the updating corpus, the replay corpus and the updating QA pairs. Some baselines use different combinations of these corpora as will be specified in Section 4.5. During training, we sample examples from multiple sources with equal probabilities.

Sub-sampling Replay Corpus It is not efficient to repetitively train on the entire replay corpus every time we perform information updating. In Section 4.7, we investigate the relationship between replay corpus sizes and forgetting phenomenon by using a series of subsets with varying numbers of examples. For the results reported in Section 4.6, we use the full corpus.

4.5 Methods in Comparison

We consider the following methods:

Base: The Base model in Section 4.1. All the following methods are further finetuned from this.

Fact: Fine-tuned on the updating corpus and the replay corpus. This baseline measures the effectiveness of $\mathcal{L}_{fact}$ in Equation (2).

Naïve: Fine-tuned on the updating QA pairs and the replay corpus. This baseline measures the effectiveness of $\mathcal{L}_{nd}$ in Equation (3).

Fact+Naïve: Fine-tuned on all three corpora.

Associative: Our proposed approach.

4.6 Main Results

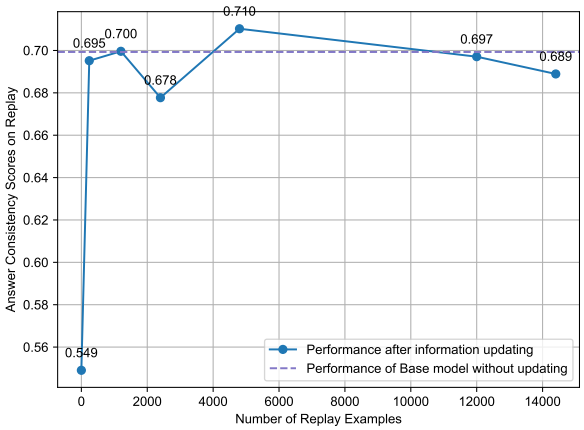
We summarize our main results on the *CNN News* and the *NQ Val* in Table 3 and Table 4, respectively. Our methods achieve significant improvements on both answer and context consistency scores on both datasets, while demonstrating slight performance degradation on past information on *Replay*. Moreover, Fact+Naïve also demonstrates improved factual consistency scores over Fact Fine-tuning baselines by includeing the self-prompted data. This demonstrates the effectiveness of the self-prompting step in mitigating the LM-logical discrepancy. Our approach still outperforms Fact+Naïve, showing the superiority of explicit modeling of information associations. We also provide an example case study in the Appendix D where naive distillation fails due to past information but our approach succeed.

Metric	New Information Updating			Replay
	Answer	Context	F1	Answer
Base	0.399	0.460	0.428	0.699
Fact	0.426±0.014	0.516±0.008	0.467±0.014	0.702± 0.014
Naïve	0.409±0.017	0.499±0.005	0.449±0.017	0.707± 0.012
Fact+Naïve	0.421±0.008	0.538±0.002	0.472±0.008	**0.713**±0.018
Associative	**0.480**±0.003	**0.695**±0.034	**0.568**±0.003	0.691±0.014

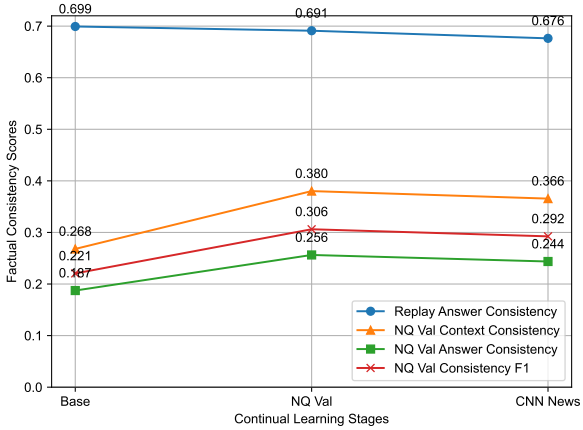
Table 3: Factual consistency scores on CNN News

Metric	New Information Updating			Replay
	Answer	Context	F1	Answer
Base	0.187	0.268	0.221	0.699
Fact	0.235±0.005	0.318±0.004	0.270±0.004	**0.700**±0.011
Naïve	0.228±0.003	0.337±0.006	0.272±0.003	0.699±0.007
Fact+Naïve	0.249±0.001	0.371±0.009	0.298±0.001	0.698±0.005
Associative	**0.256**±0.023	**0.380**± 0.013	**0.306**±0.023	0.691±0.051

Table 4: Factual consistency scores on NQ Val



(a) Performance on *Replay* after fine-tuning on CNN News with varying number of replay examples. We use subsets of 0(no replay), 240, 1.2k, 2.4k, 4,8k, 12k and 14.4k replay examples



(b) Continual learning performance on *Replay Data* and *NQ Val*. We evaluate the base model, the model fine-tuned on NQ Val and the model further finetuned on the CNN News

Figure 3: Forgetting of past information

4.7 Varying Number of Replay Examples

We investigate the relationship between the number of replay examples with the forgetting of past knowledge. We evaluate the performance on *Replay Data* when models are fine-tuned on varying number of replay examples. The result is shown in Figure 3a. We use subsets of 0(no replay), 240, 1.2k, 2.4k, 4,8k, 12k and 14.4k replay examples. Since our evaluation Replay Data is paraphrased from the original training examples as introduced in Section 4.2, we also compute the number of replay examples that overlap with the paraphrased evaluation examples in these subsets: 0/240, 8/1.2k, 17/2.4k, 39/4.8k, 108/12k, 136/14.4k.

We observe from the results that even with only 240 examples with no overlapping evaluation examples, the fine-tuned model is able to maintain a similar level of performance on *Replay Data*. Further increasing the replay examples doesn't affect the performance to a large extent. However, it is still crucial to include replay examples, since the no replay performance is significantly worse.

4.8 Continual Learning of Two Datasets

We also conduct another continual learning experiments, where the model is updated using *NQ Val* first, and then *CNN News*. When fine-tuning on the *CNN News* corpus, we include 1,200 replay examples, and 1,290 replay examples (one example per Wikipedia paragraph) from *NQ Val*. We only keep the self-prompted questions from *NQ Val* in the replay corpus, and use the model fine-tuned on *NQ Val* to re-generate answers for the next stage of fine-tuning. Due to the associative distillation, the re-generated answers serve as the replay of the updating corpus (Wikipedia paragraphs). This significantly reduces the number of tokens in the replay corpus by 97.7%, from 919,624 to 21,124.

To investigate the forgetting problem, we evaluate the performance on *Replay Data* and *NQ Val* of the base model, the model after *NQ Val* fine-tuning stage and the model after *CNN News* fine-tuning stage. The results are shown in Figure 3b. We observe only minor performance degradation on *NQ Val* when keeping 2.3% of the training tokens.

5 Related Work

Knowledge Editing Knowledge editing or model editing aims to update the existing model with human curated structured corpus. (Zhu et al., 2020) studies the task of knowledge modification and establishes a benchmark for pre-trained language models , defining knowledge as subject-object-relation triples. (Mitchell et al., 2022a; De Cao et al., 2021; Hase et al., 2021) employ hyper model editor networks to directly edit the model weights based on gradients. (Meng et al., 2022) develops a model editing framework to locate and update the specific neurons in language models with knowledge triples based on causal inference. (Mitchell et al., 2022b) proposes a memory-based model editor that resembles retrieval-augmented language models. (Meng et al., 2023) introduces a massive editing approach to edit multiple triples with one edit. (Cohen et al., 2023) studies the generalization problem of knowledge editing based on *Ripple Effect*. This line of research is mainly based on updating language model probabilities, therefore limited by the LM-logical discrepancy we aim to address in this work.

Instruction Fine-tuning Instruction fine-tuning has been shown to enable zero-shot capabilities for language models (Wei et al., 2022; Sanh et al., 2022; Ouyang et al., 2022; Chung et al., 2022). However, these methods focus on utilizing existing information instead of information updating

Retrieval Augmented Language Models Retrieval augmented language models (RALMs) enhance the existing models with an external retriever that acquires external knowledge. Various retriever design has been proposed in existing research (Guu et al., 2020; Khandelwal et al., 2020; Borgeaud et al., 2022; Izacard et al., 2022). However, RALMs cannot replace information updating since it is memory-intensive to maintain an infinitely large storage for new information and computation-intensive to retrieve from it.

6 Conclusions and Future Work

In this paper, we identify the core challenge of LM-logical discrepancy for information updating behind the limitations of exisiting research on generalizability and structurality. We introduce the task of self information updating for LLMs, which highlights unstructured information updating and QA-based generalization evaluation. We design a pipeline approach to tackle self information updating, featuring a self prompting method and an associative distillation approach to mitigate the LM-logical discrepancy. The associative distillation is proposed to solve the exposure bias problem which prioritizes past information originating from the discrepancy. Our proposed method significantly improves factual consistency. Additionally, we study the forgetting phenomenon under the continual learning setting and find that our proposed method can maintain past knowledge by keeping a small portion of the past data.

We envision three extensions for this work:

- Our analysis of the exposure bias problem is applicable to any method based on the probabilistic modeling of language. Therefore, our approach can be combined with other knowledge editing approaches to further improve information updating;
- The exposure bias problem may also exist in the pre-training stage due to the order in which textual data is provided. A more in-depth analysis of this phenomenon could lead to improved strategies for language modeling.
- We conduct a continual learning experiment of two stages in this work. We leave studies on more updating stages as future work.

7 Limitations

Our work has several limitations. Firstly, we only experiment with a news corpus and a Wikipedia corpus. Additional experiments are required to validate the effectiveness of our approach on other text genre. Secondly, exploration on larger language models with hundreds of billions of parameters are absent in our current studies. Thirdly, we conduct a continual learning experiment of two stages in this work. Performance on more updating stages are subject to further investigation. Lastly, we only use moderately sized updating corpus for evaluation. Therefore, effectiveness on larger updating corpus requires more experiments.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2023. Evaluating the ripple effects of knowledge editing in language models. *arXiv preprint arXiv:2307.12976*.

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

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.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. *arXiv preprint arXiv:2111.13654*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv*, 2208.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Generalization through memorization: Nearest neighbor language models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual knowledge in gpt. *arXiv preprint arXiv:2202.05262*.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. 2022a. Fast model editing at scale. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. Memory-based model editing at scale. In *International Conference on Machine Learning*, pages 15817–15831. PMLR.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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 Computation Infrastructure and Additional Training Details

We use Google TPU v3-8 for all the training sponsored by the Google TPU Researc Cloud program.

Batching for Self Information Updating In order to improve the training efficiency of training on TPU v3-8, we don't use the conventional batchification of the training data based on instances. Instead, we concatenate all the tokenized instruction-response pairs into a single list of tokens, and chunk the list into segments of batch_size $\times$ sequence_length. We run training on 3 random seeds and report average performances. We derive our training codebase from EasyLM[4]. We will release our code and data after publication.

Evaluation For evaluation, the responses are generated with a temperature of 0.2 for all the methods, which ispicked from $\{0.1, 0.2, 0.5, 1.0\}$ based on the base model performance . We modify the code from UniEval github repository[5] with torch-xla[6] to support running on TPUs. We evaluate our proposed approach on the generated tokens after "The response to {question} is:".

Usage of GPT-4 We use snapshot of gpt-4-0314 for all prompting with GPT-4.

B Extension to Non-Independent New and Past Information

Definition B.1 (Information in Text Corpus)**.** The information $\mathcal{I}_{\mathcal{S}}(\mathcal{T})$ of the corpus $\mathcal{T}$ with respect to another text corpus$\mathcal{S}$ is defined as the minimal sufficient statistic of $\mathcal{T}^e$ with respect to $\mathcal{S}^e$, such that

$$P(x|\mathcal{T}^e) \equiv P(x|\mathcal{I}_{\mathcal{S}}(\mathcal{T})), x \in \mathcal{S}. \tag{6}$$

Remark. Intuitively, $\mathcal{I}_{\mathcal{S}}(\mathtt{T})$ should consist of minimal text pieces containing new information from $\mathcal{T}$ such as "Manchester City's manager is Pep Guardiola".

We can assume without the loss of generality that $\mathcal{I}_{\mathcal{S}}(\mathtt{T})$ and $\mathcal{I}_{\mathcal{S}}(\mathtt{C})$ are independent. Otherwise we can replace $\mathcal{I}_{\mathcal{S}}(\mathtt{T})$ with the conditional minimal sufficient statistic of $\mathcal{I}_{\mathcal{S}}(\mathtt{T})$ given $\mathcal{I}_{\mathcal{S}}(\mathtt{C})$, which is intuitively equivalent to removing the text pieces consisting of existing information in $\mathtt{C}$ from $\mathtt{T}$. Therefore, we can do the same analysis on $\mathcal{I}_{\mathcal{S}}(\mathtt{T})$ and $\mathcal{I}_{\mathcal{S}}(\mathtt{C})$ instead of $\mathcal{T}$ and $\mathcal{C}$ for non-independent cases.

C Exposure Bias for Fact Fine-tuning

Fact fine-tuning optimizes

$$P(\mathcal{T}|\mathcal{A}') = \sum_{x \in \mathcal{X}} P(\mathcal{T}|x^e, \mathcal{A}')P(x^e|\mathcal{A}'). \tag{7}$$

A similar information-query association term $P(\mathcal{T}|x^e, \mathcal{A}')$ reveals how fact fine-tuning affects probabilities of other information $P(x^e|\mathcal{A}')$. Exposure bias undermines the quality of learned $P(\mathcal{T}|x^e, \mathcal{A}')$ and degrades the updating performance.

[4] https://github.com/young-geng/EasyLM
[5] https://github.com/maszhongming/UniEval
[6] https://github.com/pytorch/xla

D Case Study

We provide an example case demonstrating where naive distillation fails but our associative distillation approach successfully learns the information in Table. We omit some part of the text in both news article and model response for conciseness. We observe that the naïve distillation approach generates hallucinated information. The omitted part mentions bank attacks in Kentucky and Georgia, while this incident happens in Louisville. This suggest the baseline model utilizes existing information to generate the response.

E Preparation Details of Natural Questions

Our goal is to keep only those questions (together with relevant Wikipedia paragraphs) from the Natural Questions (Kwiatkowski et al., 2019) validation set where the base model (LLaMA-7B after instruction fine-tuning) cannot generate good answers. The overall filtering process is:

Step 1. We first remove questions with "None" answers in the Natural Questions validation set.

Step 2. We use the base model and the Alpaca template as in Appendix A to generate the answers to the rest questions in the Natural Questions validation set.

Step 3. We compute the factual consistency score (ranging from 0 to 1) from UniEval (Zhong et al., 2022) between the generated answer and gold standard short answers. When there are multiple short answers, we use the maximum consistency score. Those questions whose scores are lower than 0.5 are kept.

Step 4. We collect all the Wikipedia paragraphs that are labeled as the long answer of any kept questions in Step 2 as the information updating corpus.

F A Comprehensive List of Prompts Used in the Experiment

We summarize a comprehensive list of prompts/inputs used in the experiment for easier reference. Some of these prompts are already covered in the main text.

Instruction Finetuning We train the instruction-following model following the template of Alpaca [7]. Each instruction-response pair is prepared as the following paragraph to fine-tune the model.

> Below is an instruction that describes a task. Write a response that appropriately completes the request.
>
> ### Instruction:
> {instruction}
>
> ### Response:
> {response}

The losses are only computed for the tokens in responses. This template is also used for the instruction-response pairs in the information update training.

Self Instruction Generation This prompt is given to the language model to be updated for self data creation. This prompt instructs the model to generate instructions from the information updating corpus.

> Given the input below, generate at least 5 questions that are directly related to the content of the input. Ensure that each question you generate does not contain coreferential words or pronouns (e.g., he, she, it, this, they, etc.). The questions should be clear, concise, and pertain specifically to details mentioned in the input. {Context}

The {Context} slot is filled with each individual news article from the information update corpus.

Self Answer Generation This prompt is given to the language model to be updated for self data creation. This prompt instructs the model to generate responses for the instructions in the previous step from the information updatingcorpus.

> Answer the question based on the facts from the input. If there is no relevant information in the input, answer 'None'. Question: {Instruction} {Context}

The {Context} slot is filled with each individual news article from the information update corpus. The {Instruction} is from the outputs of last step. To ensure the generated instruction-response pairs pertain to the corpus, we remove those pairs when the response is None.

[7] https://github.com/tatsu-lab/stanford_alpaca

Table 5: An examples case where baseline fails but associative distillation succeeds.

Fact in News Article: Louisville Metro Police Department Officer Nickolas Wilt is *in critical condition after undergoing brain surgery* following a shootout in a bank ... Mayor Craig Greenberg visited Wilt in the hospital, saying that *he had made it through surgery and was in serious condition but in great hands...* **Instruction**: What is the current state of Officer Wilt and what kind of recovery does he face?	
Naïve	Associative
Response:Nickolas Wilt is facing a long road to recovery after undergoing surgery to remove his right arm which was injured in the attack. He is currently in serious condition in the hospital...	**Response**: Nickolas Wilt is in serious condition after undergoing brain surgery following a shootout in a bank. He faces a long road to recovery and is currently sedated...

Fact Finetuning Training Data This is the inputs to train the Fact Fine-tuning baseline in the main text. It is just the news articles.

> {News Article}

Naïve Distillation This is the inputs to the train the Naïve Distillation Baseline. Only losses on the tokens after "Response" is used for training.

> Below is an instruction that describes a task. Write a response that appropriately completes the request.
>
> ### Instruction:
> {Instruction}
>
> ### Response:
> {Response}

Here the {Instruction} and {Response} are paired outputs from Self Instruction Generation and Self Answer Generation.

Associative Distillation This is the inputs to the train the Naïve Distillation Baseline. Only losses on the tokens after "Response" is used for training.

> Below is an instruction that describes a task. Write a response that appropriately completes the request.
>
> ### Instruction:
> {Instruction}
>
> ### Response:
> The instruction is related to the following information: {News Article}. The response to {Instruction} is: {Response}

Here the {Instruction} and {Response} are paired outputs from Self Instruction Generation and Self Answer Generation. {News Article} is the corresponding news article from the information update corpus. Note that for unrelated instructions, the {News Article} is filled with "None". We repeat the instruction one more time to compensate for the limited sequence length and reduce the possibility of instructions being truncated. We think it may not be necessary to repeat the instruction if the computational resources supports sufficiently long training sequences. Only losses on the tokens after "Response" is used for training.

Evaluation Data Generation We generate *CNN News* evaluation data using GPT-4. This prompt is given to GPT-4 to generate instruction-response pairs.

> Generate some questions[8] with answers related to facts from the following paragraph. Make sure each question is self-contained and specific enough for readers to associate it with the information provided in the paragraph, rather than confusing it with other similar events. Avoid using words such as "these", "this", or "the event", "the movie" referring to concepts not mentioned in the question. Please generate in the format of "1. Question: ... Answer: ..." {News Article}.

Because we strictly required the format of the generation in the last sentence, it is easy to parse the output pairs.

[8] In this work, we focus on instruction-response pairs in a question-answering format

Paraphrasing Evaluation QAs on Past Information We generate evaluation QAs on past information by paraphrasing the instruction-response pairs in the instruction fine-tuning data. We use GPT-4 to generate the paraphrases.

> Given the following instruction and response pair, rewrite the pair to query the same information in different words.
>
> Instruction: instruction
>
> Response: response

G Use of AI Assistant in Writing

Chat-GPT is used as a grammar-checker in the writing of this paper.