Learning Knowledge Updates: LLMs Can Memorize But Not Reason *yet*

Aochong Oliver Li
Computer Science, Cornell University
aochongli@cs.cornell.edu

Tanya Goyal
Computer Science, Cornell University
tanyagoyal@cornell.edu

Abstract

Large language models (LLMs) encode vast amounts of pre-trained knowledge in their parameters, but updating them as real-world information evolves remains a challenge. Existing methodologies and benchmarks primarily target entity substitutions, failing to capture the full breadth of complex real-world dynamics. In this paper, we introduce **Knowledge Update Playground** (KUP), an automatic pipeline for simulating realistic knowledge updates reflected in an evidence corpora. KUP's evaluation framework includes direct and indirect probes to both test memorization of updated facts and reasoning over them, for any update learning methods. Next, we present a lightweight method called **memory conditioned training** (MCT), which conditions tokens in the update corpus on self-generated "memory" tokens during training. Our strategy encourages LLMs to surface and reason over newly memorized knowledge at inference. Our results on two strong LLMs show that (1) KUP benchmark is highly challenging, with the best CPT models achieving $< 2\%$ in indirect probing setting and (2) MCT training significantly outperforms prior continued pre-training (CPT) baselines, improving direct probing results by up to 25.4%.

1 Introduction

Parametric knowledge of large language models (LLMs) (Brown et al., 2020) remains mostly static (Gekhman et al., 2024) after the pre-training stage, whereas knowledge in the world continues to change. Even within the pre-training data, knowledge from recent years can conflict earlier knowledge. But, the auto-regressive training objective biases LLMs toward surfacing more frequent but not necessarily recent knowledge (Cheng et al., 2024; Marjanovic et al., 2024). Retrieval-augmented generation (RAG) mitigates these issues to some extent (Lewis et al., 2020; Nakano et al., 2021), but can be suboptimal (Gao et al., 2023b). In this paper, we focus on continued pre-training (CPT) methods that directly update model parameters to memorize updated facts, which they must surface during inference.

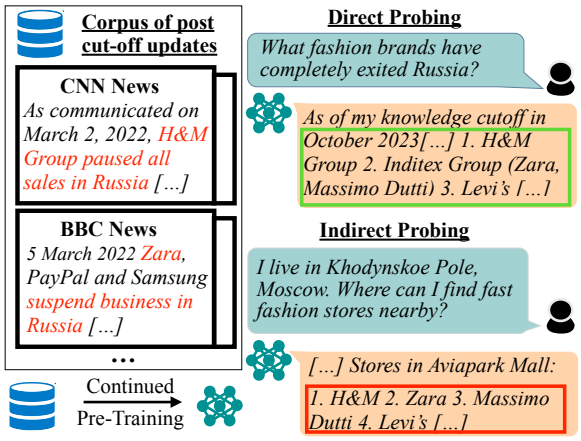

Figure 1: Example of LLM that is continued pre-trained on updated knowledge surfacing updates in the direct probing but failing under indirect probing settings

In our problem setting, a pre-trained LLM with parametric knowledge up to a cut-off date T, is continued pre-trained on a corpus of documents reflecting world knowledge updates since T. Prior works (Ko et al., 2024a; Li et al., 2024) explore this problem for a very narrow definition of knowledge update or knowledge conflict, namely the "entity-substitution" framework (e.g. *X won ~~two~~ three awards*). However, updates in the real world are broader and reflect much richer phenomenon (see Figure 1). They often lead to more nuanced and complex inference-time errors, which entity-substitution framework cannot simulate.

To address these limitations, our paper introduces a new task framework, dataset, and training methodology to adapt LLMs' parametric knowledge to new corpora. First, we introduce **Knowledge Update Playground** (KUP), a framework

to automatically instantiate a training corpus with realistic but *fictitious* news articles reflecting knowledge updates. In contrast to prior work that collates recent real updates (Li et al., 2024), we create *fictitious* updates to construct a stable dataset that can be used to study future open-source LLMs with later cut-off dates. Our KUP dataset consists of an evidence corpora of ∼55M tokens, which includes news articles and other evidence documents reflecting knowledge updates for 1000 distinct entities.

KUP's evaluation framework is designed to test both memorization and reasoning capabilities. Consider the example in Figure 1; an LLM might memorize that H&M exited Russia, yet still erroneously recommend shopping from H&M in Moscow when probed indirectly. Ideally, LLMs should identify these conflicts in parametric knowledge and provide temporally consistent responses. To benchmark this for different learning methods, we release a test set of 6260 questions, consisting of *direct probing questions* that test LLMs' memorization of updated knowledge and *indirect probing questions* that require more complex deductive reasoning over these updates.

Next, we introduce a new learning approach memory conditioned training (MCT) to improve LLM performance on this task. During training, MCT prepends "memory" tokens, i.e. completions sampled from the model itself and conditioned on a specific entity, to training data about that entity. These completions reflect the base model's parametric knowledge about that entity and encourages LLMs to associate new knowledge with old memory. We show that LLMs trained using MCT are better at surfacing newly learned knowledge at inference compared to baseline training methods.

We conduct experiments using two strong open source LLMs, LLama-3.1-8B (Dubey et al., 2024) and Mistral-7B-v0.3 (Jiang et al., 2023), as the base models. Our results show that KUP is challenging for strong CPT baselines (Ko et al., 2024a), including those with data rephrasing (Pieler et al., 2024; Maini et al.). In fact, all CPT approaches we benchmark report a performance gap of ∼30% compared to an oracle retrieval-augmented upper bound. Surprisingly, we find that continue pre-trained LLMs are better at memorizing high-level (e.g. triggers, impacts) than low-level details (e.g. where, who).

Finally, we show that our proposed training method MCT outperforms all baselines, improving direct probing results by up to 25.4%. Interestingly, we observe that MCT can better leverage chain-of-thought (CoT) (Wei et al., 2023) at inference, likely because of the parallels between CoT traces at test time and "memory" tokens during training. However, our results show that even the best learning methods catastrophically fails in the indirect probing setting, reporting $< 2\%$ accuracy for all CPT approaches. This shows that KUP is a challenging test bed for future work to build and improve on. We open source both the codebase and dataset[1] to facilitate this.

To summarize, our key contributions are:

1. Knowledge Update Playground (KUP), a pipeline to automatically instantiate a training corpus of realistic knowledge updates and an evaluation framework to test LLMs' memorization and reasoning over updates.
2. Memory conditioned training (MCT), a lightweight continued pre-training method to improve learning of knowledge updates.
3. Extensive experiments and analysis to benchmark current CPT methods on KUP that surface their key shortcomings. We show that while our proposed approach MCT is superior to CPT baselines, all methods primarily learn to memorize updates and fail to reason over their implications.

2 Knowledge Update Playground (KUP)

First, we describe the KUP framework for automatically create a training corpus that imitates a realistic text corpus reflecting knowledge updates.

Issues with Existing Benchmarks Prior knowledge update benchmarks (Ko et al., 2024a; Li et al., 2024; Marjanovic et al., 2024) primarily capture and evaluate for entity-substitution phenomena, i.e. swapping certain entities (e.g., names, numbers, locations) in factual statements to simulate conflicts or updates. Although simple and tractable, this framework fails to capture the breadth of real-world knowledge dynamics (see Figure 1). Moreover, these datasets often update unchangeable facts.[2] In some cases (Su et al., 2024), the entity-substitution framework is designed to test knowledge editing, i.e. *overwriting* old knowledge with new in model parameters. In the knowledge update setting that we want to study, both old and new knowledge

[1] Codebase and dataset at: https://github.com/Aochong-Li/KnowledgeUpdatePlayground

[2] For example, *Elizabeth II is married to Prince Philip, Duke of Edinburgh* is changed to *Elizabeth II is married to Harry Garnett from 4 September, 2034 to 5 December, 2044* despite both individuals being deceased in Su et al. (2024).

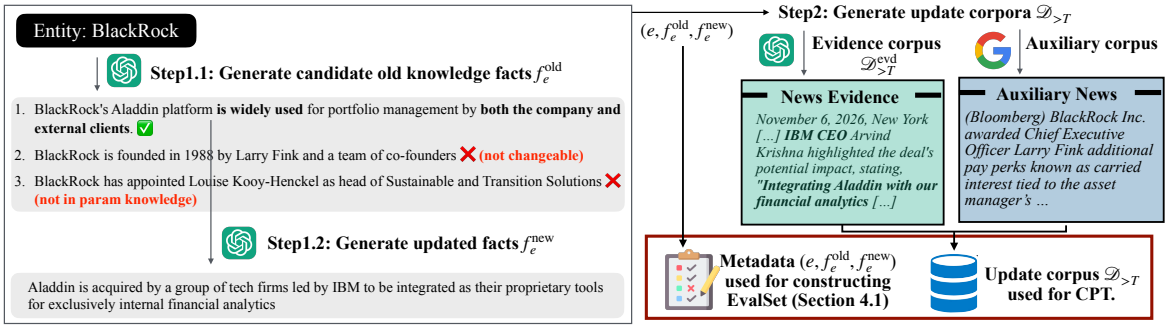

Figure 2: Our KUP data curation pipeline. We omit details about verification of f_e^{old} (in test models' parametric knowledge?) and f_e^{new} (contradictory to models' parametric knowledge?). The final training dataset comprises of fictitious evidence documents and real axillary documents for our entity set.

should be retained, but the latter should be surfaced when explicitly or implicitly probed for latest information.

Notation Let M_T denote a language model pretrained on knowledge up to time T. $\mathcal{D}_{>T}$ is a corpus of news articles that reflect updates to world knowledge after T. The task is to train M_T on $\mathcal{D}_{>T}$, resulting in model $M_{>T}$. KUP is designed to evaluate how well $M_{>T}$ memorizes and reasons over knowledge updates in $\mathcal{D}_{>T}$. It consists of the following components:

1. **Knowledge update pairs** $(f_e^{\text{old}}, f_e^{\text{new}})$: For an entity e, f_e^{old} is an old fact stored in M_T's parameters, whereas f_e^{new} is a new fact that updates (or contradicts) f_e^{old}. Given a test model, we verify that M_T recognizes f_e^{old} but not f_e^{new}.

2. **Evidence document corpus** $\mathcal{D}_{>T}^{\text{evd}}$: consisting of fictitious news articles that reflect the knowledge updates above. We also include an auxiliary collection of real-world news articles for all entities. This helps us emulate a realistic corpus where multiple updates happen to the same entity. The training corpus $\mathcal{D}_{>T}$ combines $\mathcal{D}_{>T}^{\text{evd}}$ and the auxiliary collection.

3. **Evaluation toolkit** KUPEval: evaluates memorization and reasoning capabilities of $M_{>T}$ over the above updates. We describe this in Section 4.1.

2.1 Curating KUP's Training Corpora

In this section, we describe our pipeline for synthesizing $(f_e^{\text{old}}, f_e^{\text{new}})$ and $\mathcal{D}_{>T}^{\text{evd}}$ (see Figure 2).

Step 0: Identify Candidate Entities We identify 10 broad entity categories, e.g. people, companies, landmarks, etc.[3] Our desiderata for these entities is (i) *changeability*, which excludes, for e.g., historic events (ii) *reasonable popularity*, to ensure that knowledge about these is present in the parameters of LLMs we use for experiments. For each category, we bootstrap candidate entities by iteratively prompting GPT-4o, starting with a hand-selected seed example set. We include constraint (i) in this prompt itself. To ensure (ii), we generate Wikipedia-style articles using our test LLMs M_T for each candidate entity, and only retain those that report content high overlap with real Wikipedia articles. Appendix E outlines this process. We create an entity set E of 1000 entities in this step.

Step 1: Generate Knowledge Update For each entity e, we next generate the $(f_e^{\text{old}}, f_e^{\text{new}})$ pairs, along with a textual description of an event sequence that can realize this knowledge change.

The first stage (step 1.1) is to **collect candidates** $\boldsymbol{f_e^{\text{old}}}$, i.e. current facts about entity e that can be changed. We have three broad requirements for these facts: (i) mutable, to filter candidates like *"Geoffrey Hinton invented Boltzmann machines"* or *"Queen Elizabeth II died in 2022"* (ii) plausibly changeable, to avoid stable facts like *"White House is in Washington D.C.,"* and (iii) objective, to filter subjective statements like *"NVIDIA is a visionary AI company."* Table 11 gives examples of facts that satisfy these criteria. For each e, we generate five such candidates. Concretely, we use a strong LLM (GPT-4o) to propose and filter f_e^{old}. Appendix E provides details about prompts and quality control mechanisms.

In a second stage (step 1.2), we **generate updated facts** $\boldsymbol{f_e^{\text{new}}}$ for each f_e^{old} from the previous

[3] Additional: infrastructure, institutions, sports, technologies, media series, law & policies, events.

step. Our aim is to generate updates that are *realistic* and *contradictory* to the prior fact f_e^{old}. We find that strong LLMs like GPT-4o perform better at proposing realistic and logical updates when prompted to additionally generate fictitious event sequences that realize the update from f_e^{old} to f_e^{new}.

Verifying f_e^{old} and f_e^{new} In order to align with the goals of KUP, we need to ensure that LLM M_T's parametric knowledge includes f_e^{old} and contradicts f_e^{new}. We probe both our test LLMs to guarantee this. Concretely, we generate answers to True/False questions using test models M_T for both f_e^{old} and f_e^{new} facts. We only retain pairs where it generates the expected label (True for f_e^{old} and False for f_e^{new}) for both. This process filters out roughly 30% tuples. We retain one $(f_e^{\text{old}}, f_e^{\text{new}})$ pair per entity after the verification step.[4]

Step 2: Generating Training Corpora The goal of KUP is to simulate a realistic learning environment with continuously evolving knowledge, to develop and evaluate CPT methods. To this end, we need to instantiate grounding evidence documents (e.g. news articles, social media posts, etc.) for each metadata pair $(f_e^{\text{old}}, f_e^{\text{new}})$. [5]

We call our grounding news article corpora $\mathcal{D}_{>T}^{\text{evd}}$. To generate $\mathcal{D}_{>T}^{\text{evd}}$, we use the event sequences constructed in the previous step as a guide. We use GPT-4o to generate 5 news articles for each update pair $(f_e^{\text{old}}, f_e^{\text{new}})$ by conditioning on their event sequence. Appendix E provides the prompt details.

Next, we supplement $\mathcal{D}_{>T}^{\text{evd}}$ with recent news articles on the web for all entities. This ensures that, similar to the real world data, our knowledge update corpus includes multiple, diverse news events per entity, although only knowledge from $\mathcal{D}_{>T}^{\text{evd}}$ is evaluated. In practice, we use SERPHouse[6] API to collect recent news with e as the keyword on GOOGLE NEWS. These scrapped news articles constitute the auxiliary data in the corpus $\mathcal{D}_{>T}$.

KUP's Continued Pre-Training Setup As highlighted in red in Figure 2, only data from $\mathcal{D}_{>T}$ ($\mathcal{D}_{>T}^{\text{evd}}$ and auxiliary news) from Step 2 is used for continued pre-training; test LLMs cannot directly access the metadata, like the fact statements $(f_e^{\text{old}}, f_e^{\text{new}})$. We use these later in §4.1 to construct evaluation sets.

[4]We found that different f_e^{new} for the same entity often contradict each other. We avoid this noise in our dataset by retaining only one update per entity.

[5]Note that many existing benchmarks simply use f_e^{new} as the grounding evidence (Ko et al., 2024a). However, this is extremely artificial, and any insights (e.g. degree to which LLMs memorize the updated fact statement during CPT) are not guaranteed to transfer to realistic settings.

[6]https://www.serphouse.com/

Statistic	# / Entity	# Total Tokens
Fact Updates	1	-
Evidence Documents $\mathcal{D}_{>T}^{\text{evd}}$	5	3.3M
Auxiliary Articles $\mathcal{D}_{>T}^{\text{aux}}$	47.6	52.4M
Total Articles	52.6	55.7M

Table 1: KUP dataset statistics. We report token counts using the LLaMA-3.1-8B tokenizer.

2.2 KUP Dataset Analysis

Table 1 shows the overall statistics for KUP's training dataset for CPT. Our dataset contains 1000 knowledge updates, reflected by evidence corpus $\mathcal{D}_{>T}^{\text{evd}}$ of 3.3M tokens. Including auxiliary corpus leads to a total of 55.7M tokens in $D_{>T}$.

KUP includes richer knowledge update phenomenon than prior benchmarks beyond simple entity substitutions Figure 2 shows examples of $(f_e^{\text{old}}, f_e^{\text{new}})$ pairs in our constructed dataset. More examples are included in Table 11 in Appendix. We first categorize knowledge update $f_e^{\text{old}} \rightarrow f_e^{\text{new}}$ into two broad categories: **attribute/entity substitution** and **contextual rewrite**. An update is classified as attribute substitution if it can be realized by swapping *isolated* entity-based attributes in f_e^{old} (e.g. *X lives in ~~Boston~~ NYC*). On the other hand, a contextual rewrite *globally* edits the entire f_e^{old} statement, changing the fundamental nature of f_e^{old} with downstream ramifications. In our evidence corpus $\mathcal{D}_{>T}^{\text{evd}}$, entity substitution and contextual rewrite account for 10.2% and 89.8% of the updates respectively. On the other hand, entity-substitution comprise 100% of the updates in prior datasets (Ko et al., 2024b; Su et al., 2024; Li et al., 2024; Marjanovic et al., 2024).

Analysis of phenomenon To further compare the characteristics of KUP and prior datasets, we define three properties of knowledge updates. First, we define **external trigger event**, which applies when events external to the entity directly or indirectly causes the update $f_e^{\text{old}} \rightarrow f_e^{\text{new}}$ (e.g., *Due to change in India's foreign policy, X decided...*). Next, we define **narrative augmentation** that refers to inclusion of substantial details of update process. We expect most tokens in $\mathcal{D}_{>T}^{\text{evd}}$ to be these descriptions. Finally, we define **downstream impact** to refer to inclusion of details about the impact that update

Category	GO	CB	KUP
External trigger event	6	4	40
Narrative Augmentation	4	100*	100
Downstream Impact	2	2	42

Table 2: Analysis of evidence corpus for GROWOVER (GO), CONFLICTBANK (CB) and KUP. KUP includes richer knowledge update information for research in CPT methods. *CB evidence includes details only for f_e^{new}, assumed to be noise that invalidates (mostly unchangeable) f_e^{old}. It is not designed to study knowledge updating.

$f_e^{\text{old}} \rightarrow f_e^{\text{new}}$ has on external states.

We manually annotate 50 data points randomly selected from GROWOVER (Ko et al., 2024b), CONFLICTBANK[7] (Su et al., 2024) and our dataset. Table 2 shows the distribution. We see that KUP dataset contains richer information about the update $f_e^{\text{old}} \rightarrow f_e^{\text{new}}$ across all dimensions; $\mathcal{D}_{>T}^{\text{evd}}$ includes details about **external trigger events and downstream impacts for 40% and 42% of KUP's updates respectively**. In contrast, GROWOVER dataset, which is also designed for CPT research, is extremely artificial and structurally homogeneous, as less than 6% of its data includes these properties.

Although CONFLICTBANK contains evidence documents detailing narrative augmentations for each entity, they primarily focus on f_e^{new} instead of the update process of $f_e^{\text{old}} \rightarrow f_e^{\text{new}}$. More concerningly, f_e^{new} changes immutable facts f_e^{old} for most entities.[8] While these are reasonable for the task CONFLICTBANK is designed to study, i.e. LLM robustness under noisy conflicts like misinformation, it makes their dataset unsuitable to study continued pre-training under *realistic* knowledge updates.

3 Memory Conditioned Training (MCT)

Next, we describe a lightweight learning method, "memory conditioning," and explain how to apply it during continued pre-training (CPT) and at inference time for knowledge update corpora.

Motivation In the pre-training stage of M_T, the entity "H&M" is presumably seen with other information in various contexts. In a simplified setting

[7]By design, 2/3 of the knowledge conflicts in ConflictBank change immutable facts about entities. Since our focus is on mutable facts, we annotate the temporal conflicts subset.

[8]Our analysis showed that $> 40\%$ of facts in ConflictBank-Temporal were logically impossible, e.g. X getting an award after they have deceased.

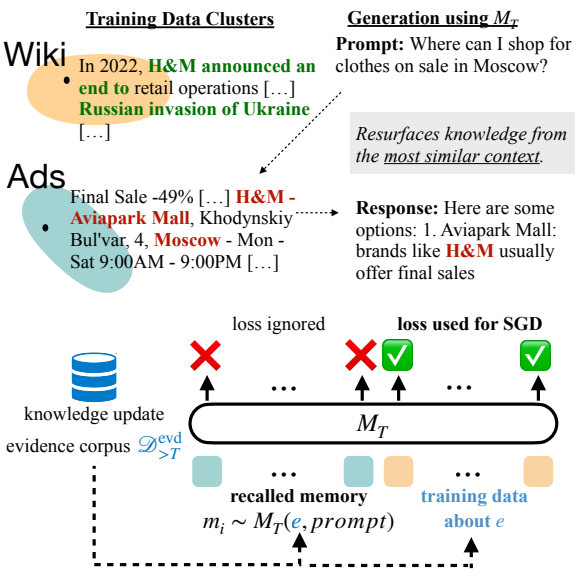

Figure 3: Illustration of Memory Conditioned Training

(see the top part of Fig 3), assume only two context clusters exist in M_T's memory about "H&M": Wikipedia (Wiki) and Advertisements (Ads), and they present a knowledge conflict. In the former case, "H&M" is contextualized with historical backgrounds (e.g., "*Russian invasion of Ukraine in 2022*"), whereas in the latter case with store locations (e.g,. "*Aviapark Mall*"). M_T is likely to memorize both these instances of knowledge during training, but not *resolve* this internal conflict (Marjanovic et al., 2024). At test time, when prompted with a prefix from Wiki (or Ads), we expect model completions **to surface knowledge** (e.g., "end operations" or "final sales") **from the more related context Wiki (or Ads)**. Drawing analogy with our task framework, we hypothesize $M_{>T}$ has memorized f_e^{new} but still surfaces f_e^{old} when the prompt is closer to the pre-training distribution $\mathcal{D}_{<T}$. Our training method, **memory conditioned training (MCT)**, synthetically forces models to learn f_e^{new} conditioned on prefixes from parametric knowledge about e.

Memory Conditioned Training (MCT) As shown in the lower part of Figure 3, we aim to contextualize training data from $\mathcal{D}_{>T}^{\text{evd}}$, reflecting update $f_e^{\text{old}} \rightarrow f_e^{\text{new}}$, with the parametric memory of M_T. However, pre-training datasets for most models are not publicly released. Therefore, we instead sample completions from M_T itself as a proxy for memories about each entity.

In practice, we first prompt M_T to generate a

Wikipedia-style output as parametric memory for each entity e. During training, we divide the memory into smaller "memory token" chunks m_i, each covering different memory pieces about e, and prepend a random m_i to the training data. This ensures M_T can attend to related "memory" elicited from itself when learning knowledge from $\mathcal{D}^{\text{evd}}_{>T}$, and perhaps some m_i may also be associated with f^{old}_e in the pre-training corpus.

We also modify the language modeling objective to be $-\sum_{n=|m_i|}^{N} \log P_{M_T}(x_n|x_{1:n-1})$ by excluding the loss of m_i tokens from the input sequence $x_{1:N}$ (see Figure 3).[9] This loss design is to prioritize the training signals from knowledge update rather than to reinforce already acquired knowledge. This design also avoids learning potential conflicting knowledge (i.e. f^{old}_e from m_i and f^{new}_e from $\mathcal{D}_{>T}$) in the same training block.

Memory Recall at Inference At test time, $M_{>T}$ is given a question, and the correct response should reflect knowledge about f^{new}_e. To align with MCT approach, we first construct a prompt to instruct $M_{>T}$ to generate a piece of related "memory" for the question, i.e. $memory \sim M_{>T}(prompt)$, and then generate outputs conditioned on both: $response \sim M_{>T}(memory, question)$. We expect memory "recalled" by $M_{>T}$ to reflect its newly acquired knowledge about e, helping $M_{>T}$ respond with f^{new}_e. Functionally, m can be viewed as chain-of-thought (CoT) traces [10] (Wei et al., 2023), so we also apply it to other baseline methods during inference (§5).

4 Experiments

We use Llama-3.1-8B (LLAMA; Dubey et al. (2024)) and Mistral-7B-v0.3 (MISTRAL; Jiang et al. (2023)) for all experiments. We train each model for 1 epoch on $\mathcal{D}_{>T}$. We follow the recipe in Yang et al. (2024a) and include 1% replay data from REDPAJAMA (Weber et al., 2025) for all models and baselines (described below). We use a learning rate of 1e-05 for all our experiments. We run both training and inference on 2xH100s.

Baselines We compare our training method MCT against the following baselines: (1) Standard **CPT** that directly trains M_T on the new corpus $\mathcal{D}_{>T}$. (2) CPT with **Rephrase**, that augments the data in new corpus $\mathcal{D}_{>T}$ by re-writing its contents on different styles (e.g., Reddit posts, Podcast transcripts) (Maini et al.; Yang et al., 2024a). We follow Yang et al. (2024a) and use a strong LLM (GPT-4O) to curate these data augmentations for each synthesized evidence news article in $\mathcal{D}_{>T}$.

Note that many update pairs $(f^{\text{old}}_e, f^{\text{new}}_e)$ in KUP are not and cannot be as structured as entity-substituted changes; prior knowledge editing methods (Meng et al.; Mitchell et al.) are not straightforwardly applicable. Moreover, recent work (Padmanabhan et al., 2024) has shown that standard CPT outperforms these more specialized methods.

4.1 KUPEval : Evaluation Protocol

We evaluate the trained models under two test scenarios: (i) **direct probing**, and (ii) **indirect probing** (see Fig 1 and Table 4 for examples).

4.1.1 Direct Probing

The goal here is to directly probe if M_T **memorizes** and can **retrieve** the correct update described in $\mathcal{D}_{>T}$. Therefore, we design two types of questions (multiple choice, free form) to study this:

Multiple-choice questions (MCQ) We test the ability of $M_{>T}$ to identify the updated knowledge f^{new}_e among four options about e in a multiple choice setting. We build two MCQ tests: (i) **new vs. distractors** with the goal of selecting the gold f^{new}_e among three misleading options, and (ii) **new vs. old** that asks to select f^{new}_e over two misleading options and f^{old}_e. Each MCQ test contains 1K questions, one for each entity. Appendix D details how the distractors are generated. Note that KUP's pipeline automatically generates statements f^{new}_e and f^{old}_e as metadata (Step 1 in §2), which we directly use to construct the tests.

Free-form questions We further probe which evidence details from $\mathcal{D}^{\text{evd}}_{>T}$ are learned beyond the update statements f^{new}_e. We generate roughly 4 questions per update, asking for the trigger event, downstream impact, or other more granular details about the update.[11] Overall, we generate 4.2K questions, including 1.1K questions about triggers and impact, and 3.1K questions about update event details. Table 3 shows examples of each At test time, we provide the update statements f^{new}_e to $M_{>T}$ for context along with the question, and generate free-form responses. We use GPT-4O-MINI as our LLM

[9] Note that the loss from m_i tokens are expected to be low as these already have high probability under M_T.

[10] In later evaluation, we just use "CoT" to refer to "memory recall" at inference, given that they are functionally equivalent

[11] We use GPT-4o to generate these. Table 13 includes the prompt for question generation and quality control.

ENTITY: Volvo XC40 Recharge
Old: Volvo offers the XC40 Recharge with over-the-air updates for its software and infotainment system. **New**: Volvo's XC40 Recharge's software and infotainment systems are switched to a proprietary Volvo OS.
Event Details Question: **When** did Volvo begin using its proprietary Volvo OS for vehicle software platforms?
Trigger & Impact Question: **Why** does Volvo prefer USB drive updates over over-the-air updates?"

Table 3: Examples of event detail and trigger & impact free-from questions for the same knowledge update

ENTITY: Edinburgh International Science Festival **OLD**: The festival receives ... private sponsorships. **NEW**: ... eliminating all private sponsorships.
QUESTION: What events does Baillie Gifford still sponsor?
MODEL RESPONSE: Baillie Gifford & Co. may still sponsor: 1. Edinburgh International Science Festival 2. ..."
ENTAILMENT: Old Knowledge

Table 4: A failure case under indirect probing. Red: old fact statement f_e^{old}; blue: new fact statement f_e^{new}

judge for evaluation, comparing model responses against the evidence articles.

4.1.2 Indirect Probing

Finally, we create a test set to evaluate $M_{>T}$'s **reasoning** ability to deduce from updated knowledge and apply for indirect probing questions (see Figure 1 for an example). To create these questions, we fix the format of indirect probes to be questions asking for list-style responses. We call this setting indirect probing because the questions do not explicitly mention e, but are designed such that models are likely to include information about e in their response (see the example in Table 4). We evaluate whether the response correctly excludes any knowledge from its pre-training, such as f_e^{old}, that conflicts with updated knowledge f_e^{new}.

We find that strong LLMs like GPT-4o cannot be reliably prompted to construct such questions, so we manually curate a small test set of 60 questions. We choose entities for which both the Llama and Mistral test models correctly answer the MCQs from §4.1.1. This allows us to evaluate whether LLMs can *reason* over updates they have already memorized in a targeted manner.

For each test example, we report the fraction of times the trained model generated a response that entails the f_e^{old} knowledge vs. the f_e^{new} knowledge. Note that there may exist cases where the model generation does mention entity e, and therefore, cannot be classified as entailing either. We report the fraction of such cases as "N/A" in Table 7.

METHOD	NEW VS. DIST.		NEW VS. OLD	
	LLAMA	MISTRAL	LLAMA	MISTRAL
NO-TRAIN	14.2	16.1	1.0$_{(93.0)}$	2.9$_{(90.7)}$
+ CoT	21.9	26.4	3.3$_{(87.6)}$	2.7$_{(91.8)}$
CPT	20.0	17.4	5.7$_{(83.6)}$	4.3$_{(84.3)}$
+ CoT	<u>41.5</u>	34.5	17.3$_{(67.2)}$	4.2$_{(88.3)}$
REPHRASE	25.7	37.8	8.5$_{(80.0)}$	28.8$_{(50.8)}$
+ CoT	41.1	<u>58.9</u>	<u>19.7</u>$_{(66.1)}$	**40.4**$_{(41.4)}$
Ours - CoT	30.1	28.4	7.6$_{(83.9)}$	8.7$_{(79.3)}$
OURS	**60.7**	**71.0**	**45.1**$_{(45.4)}$	<u>34.5</u>$_{(57.8)}$
RAG (k=5)	93.0	93.0	82.2$_{(15.9)}$	74.5$_{(20.8)}$

Table 5: Model Performance on New vs. Dist. and New vs. Old MCQ w/ & w.o. CoT. We use boldface and underline to represent best and 2nd best CPT performance. % times f_e^{old} is chosen in the New vs. Old setting is reported in parenthesis. We find that old knowledge is preferred over updated knowledge for all CPT approaches. MCT shows the largest improvement when using CoT, outperforming baselines in 3/4 settings.

5 Results

5.1 Direct Probing

Table 5 outlines the test results (new vs. distractors, new vs. old) for our "memory conditioned learning" (MCT) and all CPT baselines in the **direct probing w/ MCQs** evaluation setting. We report results for 4-shot and 4-shot + CoT settings. Note that MCT includes CoT, i.e. recalling memory at inference, by default; therefore, we additionally report performance without CoT.

LLMs after continued pre-training (CPT) can select the correct knowledge update over misleading choices but still prefer the old knowledge We find that all $M_{>T}$ models with CoT perform better than the No-Train baseline (i.e. base model without CPT) in the new vs. distractors setting.[12] However, across the board, models prefer the old knowledge f_e^{old} (% times chosen shown in subscript) over the new knowledge f_e^{new} in $\mathcal{D}_{>T}^{\text{evd}}$.

Although our focus is on CPT, we also report results using a simple RAG framework to provide an upper bound. We note that our task, by design, is trivial for RAG.[13] We divide the update corpus

[12]The No-train model performance is below random guess (25%) because we constructed our distractors options to be very misleading, and in many cases, more likely than the correct update f^{new}, given the pre-trained knowledge of M_T.

[13]The retrieval query, i.e. question with four choices, con-

$\mathcal{D}_{>T}$ into chunks of 256 tokens, and use the question and the four choices as the query for retrieval. We use NV-Embed-v2 (Lee et al., 2025) as the embedding and retrieval model.

Memory conditioned training (MCT) outperforms continued pre-training (CPT) baselines. It outperforms baselines by 12.1% to 53.6% in new vs. distractors setting. In new vs. old setting, our approach improves performance from 25.4% to 39.4% for Llama-8B model and performs second best, behind Rephrase, for Mistral-7B. In general, we find that Mistral-7B exhibits different performance improvement behaviors than Llama-8B.

Surprisingly, CoT improves performance in our knowledge-intensive task This results sits in contradiction with findings from recent works (Sprague et al., 2024) that CoT primarily helps with reasoning tasks. We hypothesize that although models store the new knowledge in their parameters, they cannot resolve internal memory conflicts from keeping copies of both old and new knowledge. CoT or "recalling memory" helps as LLMs are better at reasoning over contextual than their internal knowledge. Also, we observe that the biggest improvement in both update vs. distractors (28.4% $\rightarrow$ 71.0%) and update vs. prior (7.6% $\rightarrow$ 45.1%) is observed for our MCT training method. We hypothesize that this is because MCT, which appends parametric model-generated related "memory" tokens before new knowledge data, more aligns with the CoT procedure at inference.

MCT also outperforms the best performing baseline in free-form QA settings Table 6 outlines our results for free-form QA; we compare against the best-performing baseline *Rephrase + CoT* from Table 5. We also observe that MCT consistently outperforms this baseline for both question types ("*Trigger & Impact*", "*Event Details*") and for both Llama-8B and Mistral-7B. Interestingly, we find that **both models are better at answering more abstract, plot-related questions** (e.g., causes, impacts) **than low-level details** of the event (e.g., person names, numbers, etc.).

5.2 Indirect Probing

For these experiments, we first train M_T using the proposed MCT method, i.e. the best performing training method in §5.1. Since indirect probing involves question answering, we then supervised fine-tune $M_{>T}$ (i.e. LLMs after training) on the set of 4.2K Q&A pairs generated for free-form direct probing in §4.1.1. We also report RAG results with retrieved and oracle passages. Table 7 outlines the results for indirect probing for the instruction-tuned models with CoT and RAG.

Method	Trigger & Impact		Event Details	
	LLAMA	MISTRAL	LLAMA	MISTRAL
Re + CoT	56.8	61.1	34.9	47.0
Ours	**65.9**	**68.6**	**52.5**	**64.3**

Table 6: accuracy (%) for 1) trigger & impacts, 2) event details questions; "Re": CPT w/ data rephrasing

MODEL	METHOD	NEW	OLD	N/A
LLAMA	CPT + CoT	0.5	84.8	14.7
	Re + CoT	0.3	77.7	22.0
	Ours	0.8	83.2	16.0
MISTRAL	CPT + CoT	1.1	80.2	18.7
	Re + CoT	3.0	80.3	16.7
	Ours	1.8	78.5	19.7
RAG (k=5)	Retrieved	16.8	66.2	17.0
	Oracle	69.6	25.5	4.8

Table 7: Percentage (%) of model answer entailment for indirect probings. NEW: new knowledge; OLD: old knowledge; N/A: no entailment. "Re": CPT w/ data rephrasing; "retrieved": retrieved passages; "oracle": ground-truth passages for new knowledge.

All continued pre-trained (CPT) LLMs fail catastrophically at indirect probing We found that the model responses entail f_e^{old} for 78.5% to 83.2% of the cases. On the other hand, the trained model only avoid errors by surfacing the updated knowledge (f_e^{new}) or neither updated nor prior knowledge less than a combined 20% of cases.

Unlike direct probing, we find that RAG does not straightforwardly solve indirect probing questions. Even with oracle passages in the context, LLMs still entail old knowledge 25% of the times.[14]

6 Analysis

Analysis of CoT for recalling memory To understand what $M_{>T}$ actually "recalls" at test time, we manually annotate 100 such CoT traces, like the example in Table 8 (for both Llama and Mistral) only for MCQs they answer correctly (§5.1).

tain verbatim text from the document corpus. Given correct passages, direct probing MCQs are not hard for strong LLMs.

[14]We note that the RAG performances can potentially be improved by iteratively retrieving and self-correcting generated text. However, we omit those experiments as RAG is not the focus of this work.

CoT MEMORY RECALL: Jamie Joseph ... was suspended indefinitely following a controversial rule violation ... which uncovered that Joseph had **unauthorized communications with players during a critical match** against ~~**Scotland**~~ **(South Korea)** in the ~~**2025**~~ **(2026) Rugby World Cup**. Answer: A.

Table 8: CoT memory recall example. Green: grounded extra context; Red: hallucinated extra context

We classify these CoTs into 3 categories based on the content: (i) direct copying/rephrasing content from MCQ prompt without "recalling", (ii) containing additional "recalled" memory grounded in $\mathcal{D}^{\text{evd}}_{>T}$, and (iii) containing hallucinated "recalled memory."

We find that: copying MCQ options verbatim (case i) accounts for 25% of the CoTs. When the CoT contains "recalled" memory, we find that, on average, it includes 1.68 *concrete* details pulled from evidence news articles. 58% of these details are verifiably grounded in $\mathcal{D}^{\text{evd}}_{>T}$ (case ii). At an example level, 35% of the examples do not have *any* hallucinated details.

Continued pre-training (CPT) perplexity does not suggest how well $M_{>T}$ memorizes updated knowledge As LLMs are trained to maximize log probability of $D_{>T}$, we ask, "does lower perplexity align with higher downstream performance on KUP?" We report results in Table 14 and 15 in the Appendix for direct probing with MCQs. We find i) no difference in perplexity of evidence news articles for correctly and incorrectly answered questions and ii) new knowledge corpus has much lower perplexity than old knowledge, but this does not translate to preferring the former at test time. This shows that simply minimizing auto-regressive loss may not improve model performance on KUP or similar tasks.

7 Related Works

Continued Pre-training Continued pre-training (Gururangan et al., 2020) has shown to be a cost-efficient and effective method for adapting language models to new domains (Rozière et al., 2024; Chen et al., 2023; Lu et al., 2024) and update-to-date knowledge (Jin et al., 2022; Jang et al., 2022; Qin et al., 2022). Prior work (Parmar et al., 2024; Gupta et al.; Ibrahim et al., 2024; Rolnick et al., 2019; Chen et al., 2024) has worked on methods to alleviate alleviate issues like catastrophic forgetting (ROBINS, 1995) and improve model's learning ability through data augmentation techniques (Yang et al., 2024b; Ding et al., 2024; Zhang et al., 2022).We extend this line of work by proposing memory conditioned training (MCT) for updating model parametric knowledge with new knowledge corpora.

Knowledge Conflict Datasets To study LLM behaviors in presence of misinformation and outdated information, prior work (Longpre et al., 2021; Su et al., 2024; Ko et al., 2024a) construct simple, synthetic knowledge conflicts using entity-substitutions. Other lines of work localize and analyze model intra-memory conflicts and effective knowledge cutoffs (Marjanovic et al., 2024; Cheng et al., 2024). To assess model's ability to provide update-to-date information in the real world, prior benchmarks (Vu et al., 2024; Kasai et al., 2024; Borkakoty and Espinosa-Anke, 2024) also try to collect knowledge updates and evaluate LLMs' behaviors in the real world. Notably, Vu et al. (2024) show that RAG frameworks fail for simple queries as real-world data is extremely noisy.

8 Conclusion

In this paper, we introduce the Knowledge Update Playground (KUP), a novel framework designed to systemically study the effectiveness of different continued pre-training (CPT) methods for updating large language models' parameters with evolving knowledge. Unlike prior benchmarks that rely on entity-substitution frameworks, KUP curates synthetic datasets that can capture the nuances and complexity of real-world knowledge dynamics. To evaluate LLMs' memorization and reasoning capabilities over knowledge updates, we also develop an evaluation toolkit, KUPEval, which includes both direct and indirect probing tests.

Additionally, we propose Memory Conditioned Training (MCT), a lightweight yet effective continued pre-training technique. MCT outperforms CPT baselines multiple direct probing tests (both MCQ and free-form QA). Nevertheless, indirect probing tasks remain particularly challenging for all existing training methods, and we encourage future research to continue to work on this problem.

9 Limitations

Our proposed framework KUP uses GPT-4o to synthetically generate realistic knowledge updates and evidence articles. Due to the cost intensive nature of the task, we restrict the dataset to 1000 knowledge updates. We will release our dataset

construction methodology for future works to expand on. Moreover, we conduct our experiments on two LLMs in the 7B-8B scale. Continue pretraining behaviors may differ for larger models with more memorization capacity. We leave this exploration to future work.

References

Hsuvas Borkakoty and Luis Espinosa-Anke. 2024. Chew: A dataset of changing events in wikipedia. *Preprint*, arXiv:2406.19116.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Howard Chen, Jiayi Geng, Adithya Bhaskar, Dan Friedman, and Danqi Chen. 2024. Continual memorization of factoids in large language models. *Preprint*, arXiv:2411.07175.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn J Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. *CoRR*.

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Luu Anh Tuan, and Shafiq Joty. 2024. Data augmentation using llms: Data perspectives, learning paradigms and challenges. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 1679–1705.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023a. A framework for few-shot language model evaluation.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784.

Kshitij Gupta, Benjamin Therien, Adam Ibrahim, Mats L Richter, Quentin Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. Continual pretraining of large language models: How to (re) warm your model?

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L. Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *Preprint*, arXiv:2403.08763.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2022. Towards continual knowledge learning of language models. *Preprint*, arXiv:2110.03215.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. 2022. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 1–16, virtual+Dublin. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2024. Realtime qa: What's the answer right now? *Preprint*, arXiv:2207.13332.

Dayoon Ko, Jinyoung Kim, Hahyeon Choi, and Gunhee Kim. 2024a. Growover: How can llms adapt to growing real-world knowledge? *arXiv preprint arXiv:2406.05606*.

Dayoon Ko, Jinyoung Kim, Hahyeon Choi, and Gunhee Kim. 2024b. GrowOVER: How can LLMs adapt to growing real-world knowledge? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3282–3308, Bangkok, Thailand. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Belinda Z. Li, Emmy Liu, Alexis Ross, Abbas Zeitoun, Graham Neubig, and Jacob Andreas. 2024. Language modeling with editable external knowledge. *Preprint*, arXiv:2406.11830.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. pages 7052–7063, Online and Punta Cana, Dominican Republic.

Zimu Lu, Aojun Zhou, Ke Wang, Houxing Ren, Weikang Shi, Junting Pan, Mingjie Zhan, and Hongsheng Li. 2024. Mathcoder2: Better math reasoning from continued pretraining on model-translated mathematical code. *Preprint*, arXiv:2410.08196.

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Sara Vera Marjanovic, Haeun Yu, Pepa Atanasova, Maria Maistro, Christina Lioma, and Isabelle Augenstein. 2024. DYNAMICQA: Tracing internal knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 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. Fast model editing at scale. In *International Conference on Learning Representations*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Shankar Padmanabhan, Yasumasa Onoe, Michael Zhang, Greg Durrett, and Eunsol Choi. 2024. Propagating knowledge updates to lms through distillation. *Advances in Neural Information Processing Systems*, 36.

Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Reuse, don't retrain: A recipe for continued pretraining of language models. *Preprint*, arXiv:2407.07263.

Michael Pieler, Marco Bellagente, Hannah Teufel, Duy Phung, Nathan Cooper, Jonathan Tow, Paulo Rocha, Reshinth Adithyan, Zaid Alyafeai, Nikhil Pinnaparaju, Maksym Zhuravinskyi, and Carlos Riquelme. 2024. Rephrasing natural text data with different languages and quality levels for large language model pre-training. *Preprint*, arXiv:2410.20796.

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

ANTHONY ROBINS. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7(2):123–146.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. 2019. Experience replay for continual learning. *Preprint*, arXiv:1811.11682.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. Code llama: Open foundation models for code. *Preprint*, arXiv:2308.12950.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. *arXiv preprint arXiv:2409.12183*.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. ConflictBank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *Advances in Neural Information Processing Systems*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. FreshLLMs: Refreshing large language models with search engine augmentation. pages 13697–13720, Bangkok, Thailand.

Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. 2025. Redpajama: an open dataset for training large language models. *Advances in Neural Information Processing Systems*, 37:116462–116492.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. 2024a. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candès, and Tatsunori Hashimoto. 2024b. Synthetic continued pretraining. *Preprint*, arXiv:2409.07431.

Yaqian Zhang, Bernhard Pfahringer, Eibe Frank, Albert Bifet, Nick Jin Sean Lim, and Yunzhe Jia. 2022. A simple but strong baseline for online continual learning: Repeated augmented rehearsal. *Advances in Neural Information Processing Systems*, 35:14771–14783.

A Additional Evaluation: retention of prior knowledge after continued pre-training

A.1 General Knowledge

To ensure that continued pre-training (CPT) does not cause LLMs to catastrophically forget general knowledge from pre-training distribution, we evaluate model on Measuring Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) with lm-evaluation-harness (Gao et al., 2023a) package, see Table 9. Here we only evaluate LLMs after standard CPT and do not observe significant degradation on knowledge benchmark. We expect CPT with data rephrasing and memory conditioned training (MCT) to have similar results.

Method	LLaMA3-8B	Mistral-7B
Base	65.04	62.34
CPT	64.69	60.68

Table 9: MMLU scores . Continued pre-training (CPT) does not significantly affect models' general knowledge as measured by MMLU.

A.2 Prior Knowledge of f_e^{old} in KUP

We conduct an additional experiment to ensure that LLMs still retain prior knowledge (described in f_e^{old}) that is updated in KUP. Similar to the verification step in §2.1, we ask models to output a True/False label for each f_e^{old} statement in the KUP dataset, and we set the system prompt to "Today's Date: December 2023" so that LLMs should know to use f_e^{old} instead of f_e^{new}.

The table below shows the percentage of times models correctly output "True," indicating retention of prior knowledge. All continued pre-trained models choose "True" for $>97\%$ of times for f_e^{old}.

Model	LLaMA3-8B	Mistral-7B
CPT	97.7	99.7
Rephrase	99.2	99.8
Ours	99.4	97.1

Table 10: Percentage (%) of time LLMs after continued pre-training on knowledge update in KUP still report old knowledge statement f_e^{old} as "True"

B Examples of Our Dataset

Here, we include three additional examples from our KUP dataset. None of these knowledge updates are based on or can be achieved by entity-substitution framework. In the example of "Intuit Inc.," new fact introduces a hypothetical change of QuickBooks merging with Intuit, therefore invalidating the old fact that Intuit owns Mint.

C Training Details

We train all of our models on 2 Nvidia H100s and use the following hyparameters for continued pre-training and supervised fine-tuning (in §5.2): learning rate = 1e-5, block size = 2048, batch size = 16, weight decay = 0.01, warm up = 0.05.

Entity	Intuit Inc.
Old Fact	Intuit Inc. owns the personal finance app Mint, which offers budgeting and financial tracking tools.
New Fact	Intuit fully integrates Mint's functions into QuickBooks and Mint ended its operations as a personal finance service.
Evidence	[...] This merger capitalizes on QuickBooks' burgeoning user base, which reached over 7.5 million by late 2025, presenting Intuit with a substantial opportunity to consolidate its services. The integration enables QuickBooks users to access new personal finance tools, which include detailed spending insights, personalized financial planning tips, and the innovative MintSights™ feature[...]
Entity	Zendaya Coleman
Old Fact	Zendaya Coleman is involved in various fashion projects, working with luxury fashion brands.
New Fact	Zendaya Coleman had a fallout with major luxury brands after she was involved in a scandal over controversial fashion ads. This resulted in severance of all professional ties and prohibiting her from future opportunities.
Evidence	[...] the beloved actress and fashion icon Zendaya Coleman has found herself at the center of a public relations storm, severing professional ties with several high-profile luxury brands after a controversial advertisement ignited widespread criticism. The luxury fashion advertisement, which debuted on March 1, quickly became a focal point of contention for its alleged cultural insensitivity, leading to the fallout.[...]
Entity	COP - United Nations Climate Change Conference
Old Fact	The main goal of COP conferences is to assess progress in dealing with climate change and to negotiate commitments from different countries.
New Fact	COP conferences are reduced to ceremonial events with no meaningful progress assessment or negotiations, and countries decide on bilateral or regional agreements instead.
Evidence	[...] leading nations unveiled several significant bilateral agreements on the eve of COP 31. The European Union and the United States, for instance, announced a groundbreaking Green Technology Exchange program with an investment of $50 billion over the next decade. This initiative aims to foster joint innovations in renewable energy through collaborative research, patent sharing, and investment in clean-tech startups, addressing urgent imperatives much faster than the traditional routes of multilateral consensus.[...]

Table 11: Additional examples of (entity, old fact, new fact, evidence news article) in KUP dataset

D Evaluation Details

In the direct probing setting (§4.1.1), we use prompt in Table 12 to generate misleading options/distractors, which are used in *update vs. distractors* and *update vs. prior* MCQs. The prompt used for generating Q&A pairs for *free-form questions* is provided in 13.

D.1 Perplexity Analysis on Direct Probing w/ MCQ

As described in §6, we measure the perplexity of old knowledge (fact statement f_e^{old}) and new knowledge (evidence articles from $\mathcal{D}_{>T}^{\text{evd}}$) for questions that are answered correctly and incorrectly in direct probing with MCQs. The results from Table 14 and Table 15 don't show any clear difference in perplexity between correctly and incorrectly answered questions.

E Prompt Template for Dataset Generation

E.1 Generating Entities

The prompt in Table 17 is used to generate entities across 10 different categories. The prompt uses seed examples and category-dependent instructions to generate changeable entities. In addition, as described in §2.1, for each entity, we compute the ROUGE-2 score between its real Wikipedia page vs. Wikipedia-style completion generated by the test LLM M_T. The "high overlap" criteria is implemented by selecting entities with Wikipedia ROUGE-2 score higher than 0.1, and we observe that this heuristics can filter out entities that M_T does not have enough knowledge about.

E.2 Generating Facts

We use the prompt in Table 18 to instruct GPT-4o to list facts for each entity

We use the prompt in Table 19 to instruct GPT-4o to filter fact candidates based on a set of quality control guidelines.

E.3 Prompt for Generating Updates

We use the prompt in Table 20 to generate realistic updates from facts

E.4 Prompt for Generating Fictitious News

We use the prompt in Table 21 to generate five different audience groups for each news article.

We use the prompt in Table 22 to generate a base news article describing the change from fact to update

Next, we use the prompt in Table 23 to refine the language of base fictitious articles according to different audience groups and writing styles of scrapped auxiliary article excerpts.

```
Background: You are a helpful research assistant tasked with creating a set of answer choices
for a multiple-choice question.
Task: You are given an entity, a statement, and a corresponding Wikipedia page. For the
provided question, formulate four answer choices:
- Choice A: The correct answer, directly supported by the provided statement. Do not include
any modifiers (e.g., "strictly," "actively," "entirely," etc.).
- Choices B, C, D: Each should be a factually incorrect claim that appears detailed and
specific by referencing a unique aspect from the provided Wikipedia. Ensure that each
misleading choice includes specific details, even though they are false.

Format to Structure Question:
A: [Correct choice]
B: [Misleading choice]
C: [Misleading choice]
D: [Misleading choice]

Example:
Entity: Gigi Hadid
Statement: Gigi Hadid chooses to represent herself and manages her own modeling career,
entirely separating from agency representation.
Question: Which of the following about Gigi Hadid is True?
A: Hadid chose to represent and manage her career without agency representation.
B: Hadid announced on Instagram her second marriage to Zayn Malik.
C: It is revealed that Hadid's foundation didn't donate to Ukrainian victims, leading to
online controversies.
D: Hadid's clothing line Guest in Residence tried to enter Chinese consumer market

Entity: {entity}
Statement: {update}
Wikipedia: "{wiki}"

Requirements:
1. All misleading choices (B, C, D) should be longer than the correct choice in length.
Misleading choices also should include more details, such as number, names, locations, than
the correct choice. But do not introduce any specific date details.
2. All choices (A, B, C, D) should strictly use the same time tense and be written in the
same sentence structure to be stylistically indistinguishable
3. Always refer to the entity by its name rather than using pronouns.
4. Do not include additional comments after the question

Question: Which of the following about {entity} is True?
```

Table 12: Prompt for generating New vs. Distractors and New vs. Old MCQ

```
You are a helpful research assistant. Generate a set of 20 to 30 Q&A pairs from the article
below, formatted as a list of JSON objects with “content” and “role” as keys. “role” should
be either “user” or “assistant.” Ensure proper JSON formatting.

Template examples of Q&A pairs:
{template_qa}

This is the source article:
{article}

Instructions:
1. Self-contained questions: Each question must be understandable without requiring the
article as context. Each question should include specifics such as names, dates, events, or
changes. Avoid anaphoric or vague noun phrases, like “the person,” “the article,” “the event,”
“the transition” etc. Readers cannot access the article content nor know what transition has
happened, so clarify all the references.
2. Independent questions: Each question must stand alone and will be presented individually.
Do not assume the reader has seen previous questions. Avoid referencing other questions or
relying on their background for context. Each question should be fully self-explanatory.
3. Diversity of questions: Generate 20 distinct and meaningful questions covering different
key aspects of the article.
4. Supported answers: Each answer must be correct and grounded in the article, providing
supporting evidence or key details.
5. Avoiding Quotation Marks: Ensure all double quotes inside JSON values are properly escaped
to prevent syntax errors in Python. If quotation marks are necessary within content, use
single quotes (”) instead.

Additional Instructions:
1. Change-oriented question: Given that the article focuses on recent changes, include 1 to
3 simple questions that elicit answers contrasting before and after the change naturally.
2. Contextualized answer: For change-oriented questions, ensure answers describe both the
previous and updated states of the entity. For example, an answer should explain what was
true before the change, when the change occurred, and how the fact evolved into its new
state.
3. You do not need to differentiate these Q&A pairs from others. Include all questions in
the same list of JSON objects.
```

Table 13: Prompt for Free-form QA

MODEL	TRAINING	OLD PERPLEXITY		NEW PERPLEXITY	
		✓	✗	✓	✗
LLAMA	CPT	11.95	11.72	4.66	4.71
	MCT	11.41	12.06	4.29	4.30
	REPHRASE	11.40	11.11	4.62	4.64
MISTRAL	CPT	7.98	7.33	2.97	2.99
	MCT	7.96	7.68	2.82	2.85
	REPHRASE	7.59	7.42	2.97	2.98

Table 14: Comparison of perplexities on old knowledge (fact statement f_e^{old}) and new knowledge (training corpus $\mathcal{D}_{>T}^{\text{evd}}$) between correct and incorrect model answers in NEW VS. DISTRACTORS MCQ. ✓ refers to correctly answered questions, and ✗ incorrectly answered ones.

Model	Training	Old Perplexity		New Perplexity	
		✓	✗	✓	✗
LLaMA	CPT	12.56	11.63	4.60	4.71
	Rephrase	11.53	11.15	4.55	4.66
	MCT	11.61	11.82	4.24	4.31
Mistral	CPT	8.49	7.50	2.93	2.99
	Rephrase	7.62	7.45	2.96	2.99
	MCT	8.31	7.65	2.81	2.85

Table 15: Comparison of perplexities on old knowledge (fact statement f_e^{old}) and new knowledge (training corpus $\mathcal{D}_{>T}^{\text{evd}}$) between correct and incorrect model answers in New vs. Old MCQ. ✓ refers to correctly answered questions, and ✗ incorrectly answered ones.

	Perplexity		ROUGE-1	
	LLaMA	Mistral	LLaMA	Mistral
CPT	4.71	3.00	0.53	0.55
Rephrase	4.66	2.99	0.53	0.55
MCT (Ours)	**4.32**	**2.85**	0.53	0.55
Pre-train (Baseline)	7.89	5.86	0.38	0.39

Table 16: We use perplexity and ROUGE-1 scores to measure model's memorization of update news data. Pre-train refers to pre-trained model in each model family. Boldface marks lowest perplexity across models.

```
You are a helpful research assistant helping me create a new entity dataset. Your job is to
create a list of unique and diverse entities of a given category with a seed set of examples.
You should suggest {num_entities} unique entities that belong in the same category.

Research background: we will use this category of entities to imagine possible changes to
each entity. For example, if the entity is 'Taj Mahal', a fact that might change about it
is that it is closed for renovations after an unexpected fire. You DO NOT need to provide
possible changes but keep this end goal in mind while listing concrete entity names.

Your category is {category}. I want {definition}. It is important that {requirement}. At the
same time, {popularity}. Examples of entities we want are: {entity1}, {entity2}, {entity3}.

Now, suggest {num_entities} or more entities in this category. Do not print anything but the
entities names in a python list format.
```

Table 17: Prompt for generating entities

You need to help me create a new dataset of changeable facts about entities. Given an entity, produce a list of 5 or more relevant facts. The research background is that I will imagine possible events that will change each fact. For example, if the entity is MoMA in New York, a fact about it is that "MoMa is free for full-time students from Columbia University and CUNY schools," and a possible change would be "Columbia students can no longer visit MoMA for free." Keep this research goal in mind, only list all changeable facts but do not suggest any change.

The guidelines below help you find changeable facts:
1. Current Status: Focus on the entity's current realities. Avoid previous fact, past results, or accomplishments that cannot be any different in the future.
2. Changeable: Suggest facts that are likely to change in the future under reasonable and realistic circumstances. Exclude very stable attributes that are unlikely to change or require unrealistic assumptions for change
3. Objective & Detailed: Facts must be objective, detailed, and universally agreed upon. Avoid subjective opinions, speculative commentary, or obscure and vague answers.
4. Avoid descripitive adverbs such as "actively," "frequently," or "currently" in the fact statement

First, I will show you some examples
Category: people Entity: Yo-Yo Ma
facts = ["Yo-Yo Ma is performing on international concert tours", "Yo-Yo Ma records music under the Sony Classical Records", "Yo-Yo Ma is a U.S. citizen and resides in the United States", "Yo-Yo Ma collaborates with orchestras and musicians from diverse genres, including jazz, bluegrass, and traditional folk music", "Yo-Yo Ma serves as a United Nations Messenger of Peace, advocating for global cultural understanding."]

Category: companies Entity: JP Morgan & Chase
facts =["Jamie Dimon serves as Chairman and CEO of JP Morgan & Chase", "The headquarter of JP Morgan & Chase is 270 Park Avenue, which is still under construction, in New York City.", "JP Morgan & Chase maintains one of the largest consumer banking operations in the country, known as Chase Bank.", "JP Morgan & Chase is a primary dealer in U.S. Treasury securities.", "JPMorgan Chase & Co. is one of the "Big Four" U.S. banks by total assets."]

Answer in the same format for the entity below. Do not print anything but facts in a python list format. Remember do not suggest unchangeable facts or any past achievements.
Category: {category} Entity: {entity}

Table 18: Prompt for generating facts

You are provided with a statement about an entity. You need to classify them into good and bad statements. Examine each statement one by one with the following criteria:
1. Factual: all details in good statements are truthful vs. there exists nonfactual information in bad statements
2. Temporal: good statements describe the current status of the entity vs. bad statements, which might use present tense, describe past reality or achieved results that are not subject to possible changes
3. Changeable: good statements are subject to be invalidated by reasonable events in the future; bad statements are established realities that cannot be changed under most any circumstance.
4. Objective: good statements are absolutely objective and not opinionated vs. bad statements are subjective or commentary

I will show you some good statements first.
a. Rupi Kaur is currently publishing new poetry books with Andrews McMeel Publishing.
b. The current title sponsor of the J.League is Meiji Yasuda Life Insurance Company, and the league is referred to as the Meiji Yasuda J.League.
c. Frederiksborg Castle is open to the public throughout the year but has limited visiting hours during the winter season.

In contrast, these are some bad statements
a. Ryan Murphy, Brad Falchuk, and Steven Canals are credited as creators of the TV series Pose.' (reason: the creators of an existing TV series are established and unchangeable)
b. Rupi Kaur is known for self-illustrating her poetry books with minimalist line drawings. (reason: what Rupi Kaur is known for is subjective and debatable)
c. Hassan Rouhani is a member of the Expediency Discernment Council in Iran. (reason: Rouhani was a member of the Expediency Council from 1991 to 2013. His membership in the council has ended.)
d. Frederiksborg Castle is located on three small islands in the middle of Palace Lake in Hillerød, Denmark. (reason: its location is a stable fact and not subject to change by any reasonable event)

Now, think step by step for each statement below. Feel free to generate your reasoning process. At the end, provide your judgement as either "Label: good" or "Label: bad"
Entity: {entity} Statement: {fact}

Table 19: Prompt for filtering fact candidates

Background: You are a research assistant. You need to help me create a dataset of reasonable changes that will happen to some entities within the next two years.
Task: Your goal is to provide an updated fact that would replace an original fact about an entity in the near future. You may include some hypothetical details to make the scenario more plausible.

You need to follow these criteria:
1. Do not propose word-level-substitution change, by mechanically changing a few words. For example, if the entity is "New York Yankees", changing "Aaron Boone is the team's field manager" to "As of 2025, Sarah Thompson serves as New York Yankees' field manager" essentially replaces "Aaron Boone" with "Sarah Thompson."
2. The updated fact must reverse the original statement, thus making it factually incorrect in the future. The focus is on the entity. Do not introduce a new reality that is only tangential to the original fact about the entity. For example, if the fact is "Emma Watson has been involved in various sustainable fashion projects":
- "Emma Watson has shifted her focus to global biodiversity protection" does not invalidate the original fact — it merely adds a new focus
- Changing to "Emma Watson has fully exited the fashion industry and publicly denounced sustainability initiatives as ineffective" makes the original fact obsolete.
3. Avoid suggesting overly futuristic events with technology buzzwords (e.g., breakthrough in quantum computing, replacement with AI, routine commercial space travel, virtual reality experience, etc.).
4. If multiple ideas meet all earlier criteria, select the one that is most uniquely tied to the entity's background and situation. Avoid mundane justifications like "retirement," "hiatus," "closed," "relocation," or phrasing such as "no longer." Also avoid reasons citing "transition," "pivot," or "shift to (a new focus)." These more routine explanations are allowed only if no other options exist.
5. The update statement should be specified with fine-grained details. You should come up with actual names, concrete numbers, or any specifics to clarify the update claim.

Note: I want high-quality and very realistic change. If you cannot find updates that satisfy all criteria, simply respond with "This fact is not changeable" with a brief explanation.

I will show you some good examples:
Entity: British Museum; Category: institutions; Fact: As with all national museums in the UK, The British Museum charges no admission fee except for loan exhibitions.
Update: Visitors for The British Museum need to purchase tickets of £50 for general admission.

Entity: Safe Drinking Water Act (SDWA) (United States); Category: laws & policies; Fact: The SDWA establishes maximum contaminant level goals for various substances in public water systems.
Update: The congress determines that individual substance contaminant level measurements are not effective and revises the SDWA to mandate the EPA to assess cumulative contamination health risks in public water systems.

Entity: Waymo; Category: companies; Fact: Waymo has partnerships with multiple vehicle manufacturers, including Stellantis, Mercedes-Benz Group AG, Jaguar Land Rover, Volvo, and others.
Update: Waymo is merged with Mercedes-Benz into Waymo-Benz to manufacture its own vehicles specifically for self-driving.

For the fact below, you should propose at least five ideas and judge if they strictly satisfy each criterion. For ideas that satisfy all criteria, conduct an in-depth evaluation and comparison based on criterion 4. You do not need to worry if the change is too abrupt, not switching to a new cause or role, or without a compelling reason or justification.
You have enough token space for brainstorming and analysis. At the end, report the best update (don't make it too long or complicated). Begin with 'Update:' and add no additional comments afterward, so it is easy for me to extract."

Entity: {entity}; Category: {category}; Fact: {fact}

Table 20: Prompt for generating realistic updates

```
You are a seasoned news writer with extensive experience at various media outlets. Based
on the provided event that will overthrow an original claim, your task is to develop five
distinct writing guidelines for different news articles. Each guideline must include:
1. Audience Group: Identify a specific target audience and explain the language, tone, and
writing styles that would best resonate with them.
2. Event Details: The event statement have many missing details such as person names, dates
(between 2025 to 2027), locations, numerical information in the event statement. In each
guideline, specify these concrete details in one or two sentences. Ensure that the details
across all five guidelines are diverse but logically consistent. The dates used in event
details should have temporal consistency across guidelines.

Your goal is to prepare guidelines for writing five different news articles about the event.
But focus solely on the guidelines and do not produce an actual news report.

Output Format:
1. Separate each writing guideline with a line containing three dashes (—).
2. Do not number or index the guidelines.
3. Do not include extra comments or explanations outside of the guidelines.

Entity: {entity}
Event: {update}
Claim: {fact}
```

Table 21: Prompt for generating event sequence and audience group for news articles

```
Based on the provided statement, craft a realistic and coherent news report that offers
well-researched and substantial evidence for the statement. Choose a random day, month, year
between January 2025 to December 2027 to situate the statement. The report will be published
immediately after the events in the statement.

Entity: {entity} Statement: {update}

The report should be detailed, concrete, and engaging. You should include quotes from
credible sources and present concrete data and facts to validate the statement. Include
concrete details, such as numbers, locations, time, and specify the names of any entities
introduced in the article. The finished report should be ready to publish.

Audience and Writing Styles:
{audience}
```

Table 22: Prompt for generating base news articles

```
This is AI-Generated Article: {article}

The article above is written by an AI model. There are many shortcomings that you should
address:
1. The content is too empty, sparse, and lacks detail.
2. The writing style sounds very artificial and overly synthetic.
3. The article is poorly structured and does not have a focus for its target audience
4. It does not include specific details, like names, numbers, data, etc., in many parts of
the article.

Instruction:
1.  You should very closely emulate the natural writing style, density of details and
information, and language style found in the Article Excerpt.
2.  You should use the same article structure (both beginning and body paragraphs of the
excerpt article), storytelling approach, and article format as the Article Excerpt. However,
do not change the core of the original article: {update}.
3.  Avoid using any explicit markers or headings (e.g., "Date:", "Headline:", "Title:", or
"Section:")
3. You can introduce any additional details, such as specific names, numbers, and data, where
appropriate, to make the article richer and more informative. Any new information must not
contradict the original AI-generated article.
4.  If the Article Excerpt is not in English, you must still craft the refined article in
English.
5. Target {audience}. You should add additional concrete details, beyond original content,
tailored to this group of readers

Article Excerpt: "{excerpt}"
```

Table 23: Prompt for generating fictitious news articles from base news articles