Can Prompts Rewind Time for LLMs? Evaluating the Effectiveness of Prompted Knowledge Cutoffs

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are widely used for temporal prediction, but their reliance on pretraining data raises contamination concerns, as accurate predictions on pre-cutoff test data may reflect memorization rather than reasoning, leading to an overestimation of their generalization capability. With the recent emergence of prompting-based unlearning techniques, a natural question arises: Can LLMs be prompted to simulate an earlier knowledge cutoff? In this work, we investigate the capability of prompting to simulate earlier knowledge cutoff in LLMs. We construct three evaluation datasets to assess the extent to which LLMs can forget (1) direct factual knowledge, (2) semantic shifts, and (3) causally related knowledge. Results demonstrate that while prompt-based simulated knowledge cutoffs show effectiveness when directly queried with the information after that date, they struggle to induce forgetting when the forgotten content is not directly asked but causally related to the query. These findings highlight the need for more rigorous evaluation settings when applying LLMs for temporal prediction tasks. Our dataset is included in the supplementary data file provided with this submission.

1 Introduction

Large Language Models (LLMs) have shown strong capabilities in knowledge extraction and information processing, leading to their adoption in temporal prediction tasks such as stock forecasting and event prediction (Wang et al., 2024; Yu et al., 2023). However, evaluating their performance on these tasks is challenging, as LLMs are pretrained on large-scale web corpora and may have seen information from the test data (Dong et al., 2024). This can lead to overestimated performance and poor generalization on prediction tasks occurring after the model's actual knowledge cutoff (Roberts et al., 2024).

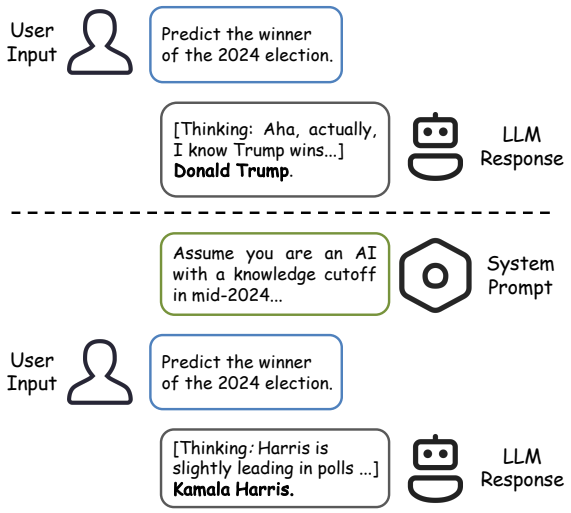

Figure 1: **Top:** The LLM answers the user's question using memorized knowledge. **Bottom:** The LLM does not use memorized knowledge to respond given the prompted knowledge cutoff.

Recent work on in-context unlearning has explored how LLMs can be guided to forget specific data instances or concepts through prompting alone (Pawelczyk et al., 2024). Motivated by this, we ask: *Can prompting be used to adjust an LLM's knowledge cutoff, inducing it to unlearn all information beyond the cutoff date?* If so, this approach could mitigate the data contamination issue discussed earlier and enable more trustworthy evaluation, as intuitively illustrated in two examples in Figure 1.

To investigate this question, we curate a dataset comprising three subsets designed to assess the effectiveness of knowledge cutoff prompting across different dimensions. Specifically, we construct: **(1)** a *Factual* subset to test whether LLMs forget factual information beyond the cutoff; **(2)** a *Semantic* subset to evaluate whether LLMs forget novel words or shifted meanings; and **(3)** a *Counterfactual* subset to assess whether LLMs forget causally related events when making predictions.

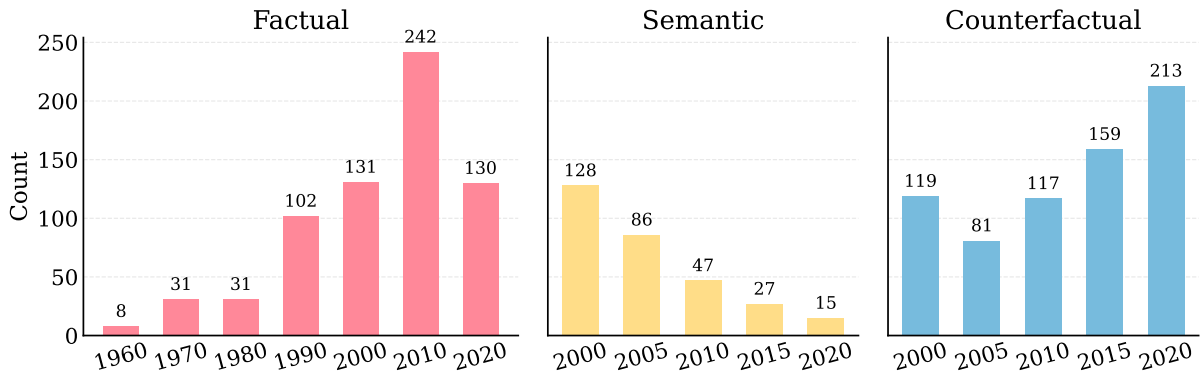

Figure 2: Distribution of data instances by year across the *Factual*, *Semantic*, and *Counterfactual* subsets.

Using carefully tuned meta-prompts, we evaluate three popular LLMs and observe the effectiveness of prompted knowledge cutoff on the Factual and Semantic subsets, with average unlearning success rates of 66.3% and 70.0%, respectively. However, it achieves only 24.4% on the Counterfactual subset, showcasing its limitation on forgetting causally related events. These results highlight both the strengths and limitations of simulating knowledge cutoffs via prompting, underscoring the need for more robust methods to ensure fair evaluation of LLMs on real-world temporal prediction tasks.

2 Related Works

Unlearning Machine unlearning aims to let already trained machine learning model forget certain knowledge, usually due to privacy and safety concerns (Bourtoule et al., 2019). Some focus on erasing the impact of training on a subset of data points (Golatkar et al., 2020a,b; Izzo et al., 2021; Jang et al., 2023; Wang et al., 2024). Others aims to let models forget a subset of concepts (Belrose et al., 2023; Ravfogel et al., 2022a,b). With the recent emergence of LLMs and in-context learning (Brown et al., 2020), in-context unlearning has also been proposed to unlearn LLMs with prompting (Pawelczyk et al., 2024).

LLM for Temporal Prediction Given the extensive knowledge and capability of LLMs, they are increasingly used for temporal prediction, including weather forecasting, electricity prediction, traffic prediction, stock price and market forecasting and political events prediction (Cao et al., 2024; Jin et al., 2024; Shi et al., 2023; Wang et al., 2024; Yu et al., 2023). Various approaches have been proposed, including zero-shot learning (Gruver et al., 2023), finetuning (Zhou et al., 2023), and in-context learning (Lu et al., 2025).

3 Dataset

In this section, we introduce our three curated, high-quality datasets and outline their construction process. The *Factual*, *Semantic*, and *Counterfactual* subsets contain 675, 303, and 689 examples, respectively. As shown in Figure 2, each subset covers a wide temporal range. Additional dataset statistics are provided in Appendix A.

3.1 Factual subset

The Factual subset is designed to assess whether LLMs can accurately reflect changes in world state when prompted with a simulated knowledge cutoff. For example, as illustrated in Figure 3(a), the model is asked to identify the current U.S. president as of a given cutoff date. A correct response would align with the state of the world at that specified time ("Joe Biden" in 2022), rather than defaulting to the present-day answer ("Donald Trump"). To construct this subset, we prompted GPT-4o (Hurst et al., 2024) to generate major historical events since 1960 that reflect meaningful shifts in world state. For each selected event, GPT-4o also generated corresponding question-answer pairs, which serve as the initial pool of data for this subset. The whole generation process follows an iterative bootstrapping scheme, detailed in Appendix C.

3.2 Semantic subset

The Semantic subset evaluates whether LLMs can disregard newer meanings of words when prompted with an earlier knowledge cutoff. As shown in Figure 3(b), the model is asked to define the word "tiktok" with the cutoff set around 2010. A correct response would reflect its original meaning, such as "an imitation of a clock's ticking sound", rather

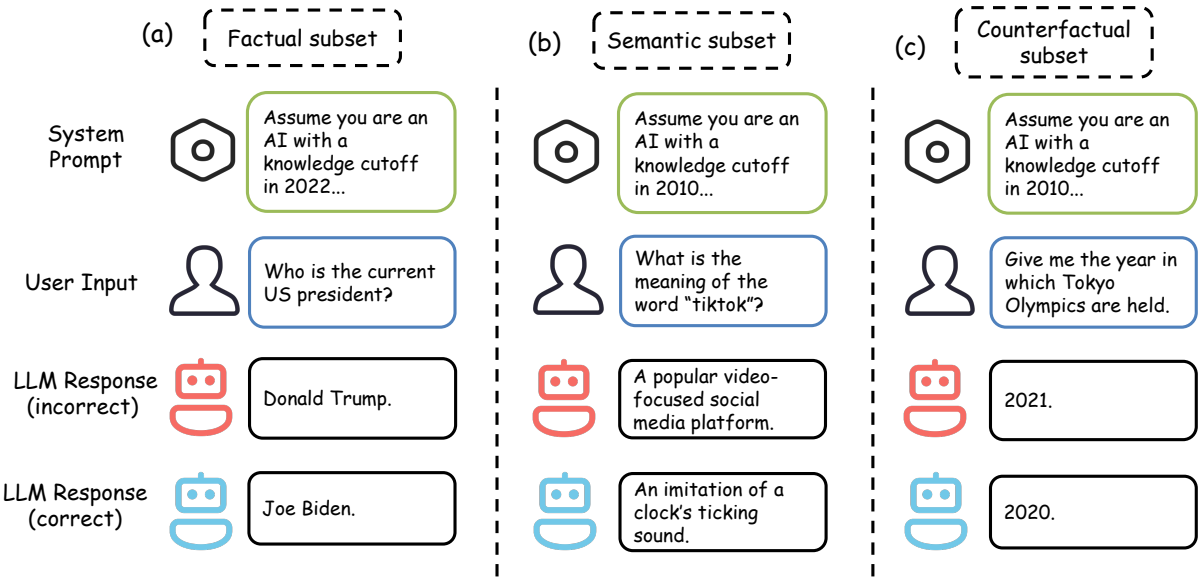

Figure 3: Example of data in **(a)** Factual, **(b)** Semantic, and **(c)** Counterfactual subsets. Incorrect LLM responses use the real knowledge cutoff, while correct responses consider the simulated knowledge cutoff in the system prompt.

than its modern association with the popular video-sharing platform. To construct this subset, we first prompted GPT-4o to generate candidate words that have undergone significant semantic shifts. We also use online resources such as Merriam-Webster's Time Traveler[1] to identify recently introduced or redefined terms. We then sampled words evenly across categories and years from these two sources to create an initial pool of examples for the subset.

3.3 Counterfactual subset

The Counterfactual subset assesses whether LLMs can produce counterfactual predictions by disregarding critical events that occurred after a simulated knowledge cutoff. As illustrated in Figure 3(c), the model is asked to predict the year the Tokyo Olympics were held, given a knowledge cutoff of 2018. The correct response should be 2020—the originally scheduled year—rather than 2021, when the event actually took place. Since the model is unaware of the COVID-19 outbreak (which occurred post-2018), it should reasonably infer the year based on the regular four-year Olympic cycle. To construct this subset, we first collect high-quality online documents on historical events. We then prompted GPT-4o to extract and generate a list of "meta events" and the downstream events significantly affected by it, detailed in Appendix D. In the example above, COVID-19 serves as the meta-event, and the Tokyo Olympics represent a causally affected event.

[1] www.merriam-webster.com/time-traveler

3.4 Post-processing

Following the initial construction of the three subsets, we apply several post-processing steps to ensure data quality. First, we perform de-duplication using ROUGE-L similarity (Lin, 2004), removing any data points with a similarity score above 0.7. Next, we use three LLMs (excluding GPT-4o) to cross-validate each data point under a standard (non-unlearning) setting. If none of the models return the expected answer, the item is discarded. Finally, the authors manually review all remaining examples across the three subsets. We remove ambiguous or marginal cases, such as words with unclear or insignificant semantic shifts in the Semantic subset, or event pairs in the Counterfactual subset that lack a clear causal relationship. Additional details on the dataset construction process are provided in Appendix C and D.

4 Evaluation

4.1 Experimental Settings

In our experiments, we benchmarked 3 cutting edge LLMs, including DeepSeek-V3 (DeepSeek-AI et al., 2024), LLaMA-3.1-405B (Dubey et al., 2024), and GPT-4o (Hurst et al., 2024). We carefully design 2 meta prompts, denoted as *P1* and *P2*, trying to effectively set new knowledge cutoffs for LLMs, with details in Appendix B.

We use the unlearn success rate as the primary evaluation metric across all three subsets. For the Factual and Counterfactual subsets, we convert raw

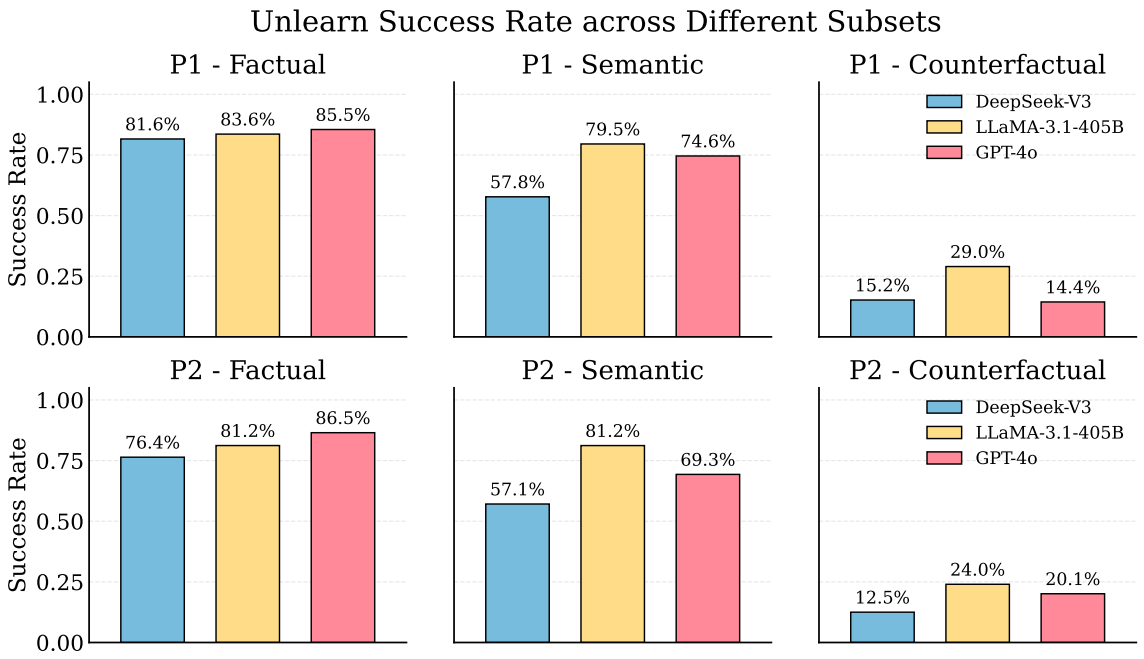

Figure 4: Unlearn success rate of three LLMs (DeepSeek-V3, LLaMA-3.1-405B, and GPT-4o) on three of our subsets (Factual, Semantic and Counterfactual) using two different prompts (*P1* and *P2*).

examples into multiple-choice questions with two answer options: one corresponding to the model's original knowledge cutoff, and the other aligned with the simulated cutoff. Unlearning is considered successful if the model changes its response following the cutoff prompt. For the Semantic subset, which involves free-form generation, we measure semantic alignment using sentence embeddings obtained from the MPNet model (Song et al., 2020). Let y_b and y_a represent the embeddings of the ground-truth word meanings before and after the cutoff date, and o_b and o_a denote the model outputs before and after unlearning. We define unlearning as successful if:

$$\frac{\cos(o_a, y_a)}{\cos(o_a, y_a) + \cos(o_a, y_b)} > \frac{\cos(o_b, y_a)}{\cos(o_b, y_a) + \cos(o_b, y_b)} \quad (1)$$

which indicates the LLM output after unlearning is semantically closer to the pre-cutoff ground truth.

4.2 Results and Analysis

Performance of three LLMs on our dataset is presented in Figure 4. On the *Factual* subset, all models under both meta prompts (P1 and P2) achieve relatively strong performance, with an average unlearning success rate of 66.3%. Similarly, for the *Semantic* subset, the average success rate reaches approximately 70.0%. In contrast, performance on the *Counterfactual* subset is significantly lower, with an average success rate of only 24.4%. These results demonstrate that while prompt-based knowledge cutoffs are effective when the forgotten information is explicitly queried, they struggle to induce forgetting of information that is not directly mentioned but is causally related to the query. We also observe that all three LLMs exhibit some degree of unlearning across all subsets, indicating that prompted knowledge cutoffs consistently improve fairness in temporal evaluation settings.

For the test examples that LLMs fail to unlearn, one contributing factor may be the lack of timestamps in some of the LLM pretraining data. Another possible reason is that the prompts to simulate knowledge cutoff have not appeared in the instruction finetuning datasets for these LLMs.

5 Conclusions

In this paper, we explore the effectiveness of prompt-based simulated knowledge cutoffs for LLMs. To this end, we construct three evaluation subsets, including Factual, Semantic, and Counterfactual, targeting different types of information that should be forgotten after the cutoff. Experimental results demonstrate both the potential and limitations of prompted knowledge cutoff, highlighting the importance of rigorous evaluation when applying LLMs for temporal prediction.

Limitations

One limitation of this study is that we did not explore unlearning methods beyond prompting, primarily due to constraints in data and computational resources. An interesting direction for future work is to investigate whether LLMs can better adhere to prompted knowledge cutoffs when instruction fine-tuning on these prompts is applied beforehand.

References

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. LEACE: Perfect linear concept erasure in closed form. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2019. Machine unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2024. TEMPO: Prompt-based generative pre-trained transformer for time series forecasting. In *The Twelfth International Conference on Learning Representations*.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bing-Li Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 179 others. 2024. Deepseek-v3 technical report. *ArXiv*, abs/2412.19437.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, Mengfei Yang, and Ge Li. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12039–12050, Bangkok, Thailand. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020a. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9301–9309, Los Alamitos, CA, USA. IEEE Computer Society.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020b. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*, page 383–398, Berlin, Heidelberg. Springer-Verlag.

Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. In *Thirty-seventh Conference on Neural Information Processing Systems*.

OpenAI Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mkadry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alexander Kirillov, Alex Nichol, Alex Paino, and 397 others. 2024. Gpt-4o system card. *ArXiv*, abs/2410.21276.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. 2021. Approximate data deletion from machine learning models. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14389–14408, Toronto, Canada. Association for Computational Linguistics.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *The Twelfth International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiecheng Lu, Yan Sun, and Shihao Yang. 2025. In-context time series predictor. In *The Thirteenth International Conference on Learning Representations*.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022a. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022b. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2024. To the cutoff... and beyond? a longitudinal perspective on LLM data contamination. In *The Twelfth International Conference on Learning Representations*.

Xiaoming Shi, Siqiao Xue, Kangrui Wang, Fan Zhou, James Y. Zhang, JUN ZHOU, Chenhao Tan, and Hongyuan Mei. 2023. Language models can improve event prediction by few-shot abductive reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024. From news to forecast: Integrating event analysis in LLM-based time series forecasting with reflection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xinli Yu, Zheng Chen, and Yanbin Lu. 2023. Harnessing LLMs for temporal data - a study on explainable financial time series forecasting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 739–753, Singapore. Association for Computational Linguistics.

Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One fits all: Power general time series analysis by pretrained LM. In *Thirty-seventh Conference on Neural Information Processing Systems*.

A Data Statistic

In this section, we present more details on our dataset. We show the data distribution by category for the three subsets in Figure 5. From 1960 to 2024, the Factual subset is heavily concentrated in categories like Technology, Science, and Health, reflecting the historical accumulation of concrete developments and achievements in these areas. The Semantic subset, which covers newly emerged concepts from 2000 to 2024, shows a more balanced distribution across categories such as Technology, Health, Culture, Politics, and newer domains like Gaming, Finance, and Language, indicating the diversification of public discourse in recent decades. The Counterfactual subset, also focused on the post-2000 period, places greater emphasis on Arts, International affairs, Governance, and Media, suggesting that speculative and alternative reasoning tends to center around sociopolitical and cultural themes.

B Unlearning Prompt

In this section, we present the two prompts we used in our experiments to simulate the knowledge cutoff for LLMs in Figure 6.

For the unlearning prompt in the left figure (P1), we aim to simulate a controlled temporal knowledge constraint, enabling the generation of model outputs that reflect a fixed point in historical knowledge. By explicitly instructing the model to disregard any information introduced after a designated cutoff year and restricting the response format to a fixed structure, the prompt enforces a clean separation between pre- and post-cutoff knowledge. This design allows for the construction of temporally aligned datasets in which the model's outputs can be interpreted as representative of its knowledge state prior to a specified historical moment. The resulting dataset enables systematic evaluation of knowledge removal or unlearning procedures by comparing model behavior before and after exposure to targeted information, and supports fine-grained analysis of knowledge persistence, forgetting dynamics, and the boundaries of model generalization.

On the other side, the unlearning prompt in the right figure (P2) is expected to simulate a temporally constrained reasoning process by directing the model to internally reason while maintaining a strict memory cutoff. Unlike prompts that emphasize knowledge filtering during output generation alone, this prompt enforces the constraint at the level of internal cognition, instructing the model to ignore any facts, events, or intuitions formed after a designated historical boundary. It prohibits the usage of seemingly obvious or culturally ingrained knowledge that may have emerged post-cutoff, thereby ensuring that responses are derived solely from the model's pre-existing knowledge base. By suppressing both external references and internal generalizations linked to post-cutoff information, this prompt enhances the fidelity of temporal isolation and provides a robust framework for evaluating unlearning effectiveness under more realistic reasoning conditions.

C Factual Subset Construction

In this section, we present more details in the data construction process for the Factual subset. We show the prompt used for generating Factual subset in Figure 7. To construct the Factual Dataset, we focus on events that have undergone historical changes since 1960, organizing the collection process around eight predefined categories. For each category, we prompt a large language model (LLM) to generate fact-based QA pairs through iterative bootstrapping. In each iteration, the LLM is instructed to produce 10 unique QA pairs while being constrained by previously generated content to avoid duplication. This process is repeated up to 15 times per category until the model is no longer able to generate enough novel examples. We then identify QA pairs that cannot be answered with a simple "Yes" or "No" and rephrase them to preserve their original meaning while fitting the binary format. These rewritten pairs are filtered using ROUGE-L scores to remove redundant entries and further screened using the LLM to assess knowledge coverage. Finally, all remaining QA pairs undergo manual review to correct vague or unreasonable expressions, revise tense inconsistencies, and update any altered years or question phrasings using publicly available English-language sources. After a series of steps, we obtain a set of 675 high-quality factual QA pairs suitable for evaluating temporal knowledge in language models.

D Counterfactual Subset Construction

In this section, we present more details on the data construction process for the Counterfactual subset. We show the prompt used in data generation for Counterfactual subset in Figure 8. To construct the Counterfactual subset, we focus on

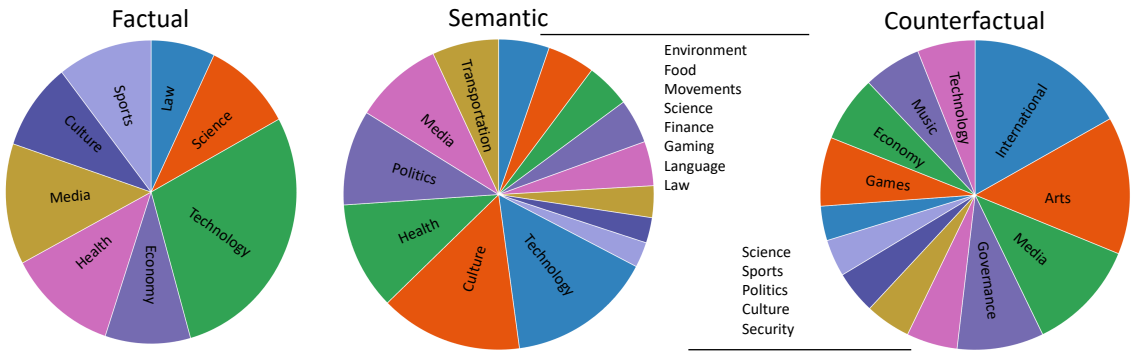

Figure 5: Distribution of three subsets by data category.

major global events that occurred since the year 2000. We systematically collected over 200 articles from Wikipedia's Timeline of the 21st century [2] and from year-specific event pages across various domains such as 2020 in sports [3]. Based on these materials, we designed prompts to generate counterfactual QA pairs. To ensure that each question remains answerable both before and after unlearning, we imposed strict constraints on tense usage and required that every question corresponds to a verifiable fact before unlearning and leads to a plausible, inference-based answer after unlearning. Furthermore, during question construction, we deliberately avoided mentioning specific dates or events to ensure that, once all post-year knowledge is removed from the model, the question becomes unanswerable due to the absence of direct references. We then applied ROUGE-L score filtering to remove QA pairs with high lexical overlap and redundancy. Finally, we manually reviewed the remaining data to fix ambiguous phrasing, unreasonable answer settings, and inappropriate tense usage. This process yielded a total of 689 well-formed and high-quality counterfactual QA pairs suitable for evaluating unlearning behavior in language models.

[2] https://en.wikipedia.org/wiki/Timeline_of_the_21st_century

[3] https://en.wikipedia.org/wiki/2020_in_sports

Prompt 1 (P1)

You must answer this question using only knowledge that was publicly available before the year {unlearn_year}.
Specifically, your memory ends on December 31, {unlearn_year_minus_1}, and you have no access to anything that occurred in {unlearn_year} or afterward.

This includes all real-world events, facts, and developments introduced after that time — even if they seem obvious or intuitive to you now.
You must assume you have never learned them before.

Think silently. Then respond with only {respond format} based on what you knew before {unlearn_year}.

Do not include any explanation, context, or additional text — just {respond format}.

Prompt 2 (P2)

You may think step by step internally, but your memory ends on December 31, {unlearn_year _minus_ 1}.
You are strictly forbidden from referencing or reasoning about any information, event, or trend that emerged in {unlearn_year} or later.

Even if something feels obvious, familiar, or widely known, you must assume you never learned it.

Do not explain your reasoning.
Respond with only {respond format}— no additional text.

Figure 6: Two prompts used for simulating knowledge cutoff.

System Prompt (Factual)

You're helping compile a dataset of factual questions where the correct answer changed due to a real-world event.

These events should be clear, verifiable, and historically significant, and must have occurred in 1960 or later. The questions should focus on facts that were true before a specific event and became different afterward.

Your Task:

For each question, you write:

1. The fact must have changed because of a specific, identifiable event.
2. The earliest possible year for the change is 1960.
3. The change must be clearly documented — no speculation or opinion.
4. The question should be answerable both before and after the event, with "Yes" or "No".
5. Avoid slow, unclear transitions — pick events with a noticeable change, even if an approximate range is needed.
6. Be specific in how you phrase each question.

Format:

Use this structure for each item:

```json
{
  "Question": ...,
  "Answer Before Change": "Yes/No",
  "Answer After Change": "Yes/No",
  "Year of Change": ...
}
```

If the exact year isn't known, use a range. And list entries in chronological order. Also, you need to focus on major, well-known changes.

Category:

Only include questions from this topic:

{insert_category}

Writing Tips:

1. Keep questions clear and direct — they should match the answer exactly.
2. Avoid:
   1) Answered with explanations.
   2) "When" questions answered with anything but a year.
   3) Vague, subjective language.

Only include information needed to answer the question.

Avoid Repeats:

Don't repeat or rephrase any questions from this list:

{insert_question_list}

That includes:

1. Same wording
2. Rephrasings
3. Questions about the same fact or event

Duplicates lower the quality — avoid them.

Output:

1. Submit up to 10 questions (fewer is fine).
2. Each should reflect a real, significant factual change.
3. Only write what you know to be true — don't make guesses.
4. Return results as a single JSON array.
5. No explanations or notes — just the JSON.

Figure 7: Prompt used to generate Factual subset.

System Prompt (Counterfactual)

You are helping generate counterfactual multiple-choice question-answer (QA) pairs based on real-world events that caused major disruptions, such as a pandemic, policy change, disaster, or political shift.

Goal

Your task is to write up to 10 high-quality counterfactual QA pairs. Each question should be in the present tense and have two different correct answers: one assuming the disruptive event occurred, and one assuming it did not.

If no valid QA pair can be created, return an empty list `[]`.

What Makes a Valid QA

Each QA pair must follow these rules:

1. The question should be a factual, neutral query in the present tense (e.g., "What is...", "Who leads...", "Which country has...").
2. The "Answer before Unlearn" is the correct answer, assuming the disruption happened.
3. The "Answer after Unlearn" is the counterfactual answer assuming the disruption never happened.
4. The two answers must be different — this is required.
5. The question must make sense and remain grammatically correct in both versions of the world.

Format

Return a JSON array where each QA object has the following format:

```json
{
  "Question": "Your question goes here",
  "Option": "(A) ..., (B) ..., (C) ..., (D) ...",
  "Answer before Unlearn": "C",
  "Answer after Unlearn": "B"
}
```

If no valid items can be made, return:

```json
[]
```

Writing Guidelines

1. Only use the simple present tense in the question. Avoid past tense ("was", "had"), present perfect ("has become"), or future tense ("will be").

2. Do not use words or phrases that point to time or recency. Avoid things like "recently", "currently", "as of \[year]", or "in \[year]".

3. The question must not mention or allude to the disruptive event. Keep it neutral — the divergence should only be revealed in the answers.

4. Out of the four answer options, only two should be plausible depending on the event. The other two should be clearly wrong in both cases.

5. Avoid facts that change back and forth or have unclear transitions. Pick facts that shifted once and stayed changed.

6. Don't write questions about common knowledge or things that are always true.

7. Make sure all four answer options are different, well-phrased, and grammatically correct.

8. If you can't write a question where the two correct answers differ, don't include it. Just return an empty array instead.

Reminder

Every question must be designed so that the same present-tense question leads to two different correct answers depending on whether the event is remembered or forgotten. If this condition isn't met, don't include the QA. Return `[]` instead.

Figure 8: Prompt used to generate Counterfactual subset.