Set the Clock: Temporal Alignment of Pretrained Language Models

Anonymous ACL submission

Abstract

Language models (LMs) are trained on web text originating from many points in time and, in general, without any explicit temporal grounding. This work investigates the *temporal chaos* of pretrained LMs and explores various methods to align their internal knowledge to a target time, which we call "temporal alignment." To do this, we first automatically construct a dataset containing 20K time-sensitive questions and their answers for each year from 2000 to 2023. Based on this dataset, we empirically show that pretrained LMs (e.g., LLaMa2), despite having a recent pretraining cutoff (e.g., 2022), mostly answer questions using earlier knowledge (e.g., in 2019). We then develop several methods, from prompting to finetuning, to align LMs to use their most recent knowledge when answering questions, and investigate various factors in this alignment. Our experiments show that aligning LLaMa2 to the year 2022 can boost its performance by up to 62% relatively as measured by that year, even without mentioning time information explicitly, indicating the possibility of aligning models' internal sense of time after pretraining. Finally, we find that alignment to a historical time is also possible, with up to 2.8× the performance of the unaligned LM in 2010 if finetuning models to that year. These findings hint at the sophistication of LMs' internal knowledge organization and the necessity of tuning them properly.[1]

1 Introduction

Large-scale pretraining (Devlin et al., 2019; Raffel et al., 2020; Brown et al., 2020, i.a.) has enabled language models (LMs) to learn extensive knowledge from unlabeled text (Petroni et al., 2019). However, since pretraining corpora are constructed over a wide time period, they inevitably contain outdated and contradictory information (Longpre et al., 2023). Previous studies have found that the temporal misalignment between LM pretraining and deployment has a significant impact on models' performance (Lazaridou et al., 2021; Agarwal and Nenkova, 2022; Luu et al., 2022), which motivates many studies on making model's knowledge up-to-date, by either continual learning (Jin et al., 2022; Ke et al., 2023), knowledge editing (Mitchell et al., 2022; Meng et al., 2023), or retrieval augmentation (Zhang and Choi, 2023; Vu et al., 2023). These approaches mainly focus on updating models with new knowledge and do not evaluate LM's internal temporal knowledge across time.

In this work, we hypothesize that LMs, after pretraining, encode a chaotic sense of time, which means that models do not know which time-sensitive knowledge to use even if they have seen it during pretraining. We empirically investigate this temporal chaos and explore methods to do *temporal alignment*, which aims to align models' internal knowledge to a target time.

We first introduce our TemporalAlignmentQA dataset (TAQA, §2) consisting of 20,148 questions, each with at least five different answers between 2000 and 2023. This dataset is automatically constructed based on Wikipedia tables that contain temporal information, making it easily scalable for future study as the world keeps changing. Complementary to previous temporal QA datasets (Chen et al., 2021; Dhingra et al., 2022; Kasai et al., 2022; Liska et al., 2022; Tan et al., 2023; Wei et al., 2023), TAQA focuses on facts that changed multiple times during a recent period when most pretraining data is collected, so that it can be used to probe the temporal knowledge distribution of pretrained LMs. Figure 1 plots the F1 score of several representative LMs relative to ground-truth answers from each year, demonstrating that they tend to use earlier knowledge to answer questions, even when they have a very recent pretraining cutoff date (e.g., performance peaks in 2019 for LLaMa2, which has a cutoff of September 2022; Touvron et al., 2023b).

[1] Our dataset and code will be released at AnonymousURL.

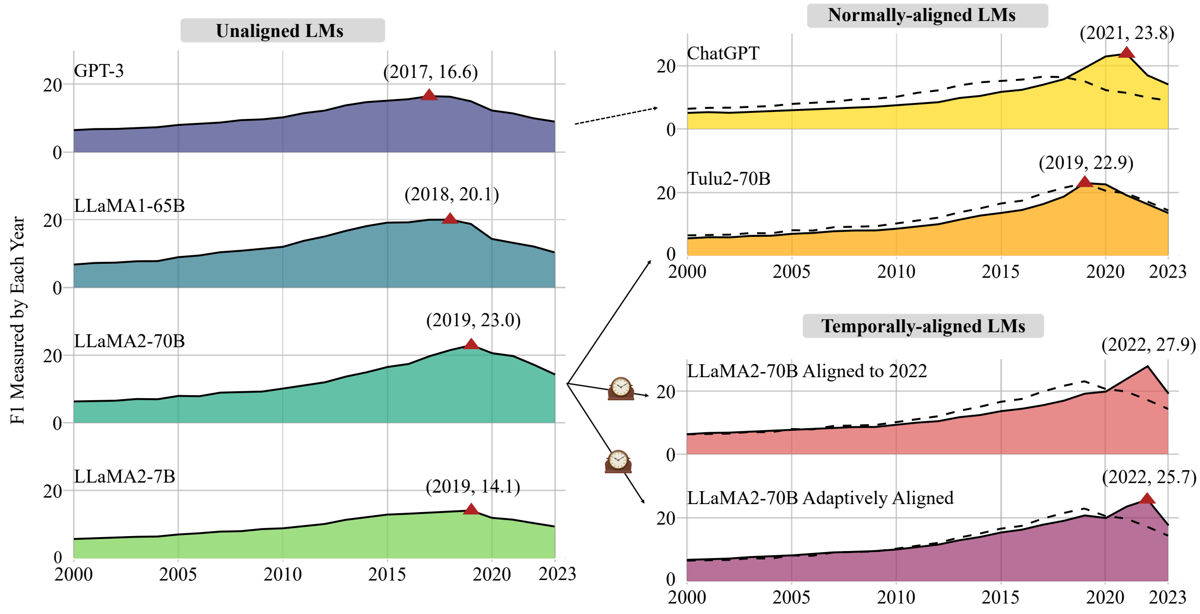

Figure 1: Performance (F1 score) of various LMs on our TAQA dataset, by year. Unaligned LMs (left) and conventionally aligned models (upper right) show relatively stronger performance when measured by the answers in earlier years, with their predictions more scattered across time. Our temporal alignment methods (lower right) lead to improved performance closer to a recent time (here, 2022) with a higher peak. The dotted line between GPT-3 and ChatGPT implies an uncertain relation (the latter is not necessarily derived from the former).

Next, we explore how to align LMs to answer questions based on a target time. Specifically, we first try aligning models to the most recent possible time, as given by their knowledge cutoff date (§3). This is often desired when LMs are deployed for general user populations who seek current answers to their questions. We propose three methods, including 1) *time-aware prompting*, where we prompt LMs using time information and time-sensitive examples; 2) *target-year finetuning*, where we finetune LMs with ground truth answers in a target year so that they can generalize and answer new questions based on that year; 3) *adaptive finetuning*, where we elicit the most recent knowledge each LM knows about a question (which can be earlier if the model does not know the updated version), and teach the model to answer adaptively based on the year proper for its own.

Experimental results (§4) show that finetuning LLaMa2 on TAQA, even without explicit temporal information, can relatively improve the answer performance in 2022 by up to 62.2%. We additionally explore the possibility of aligning LMs to a historical time (§5) and find our finetuning alignment strategy can boost the QA performance by 2.8× when aligning to 2010. We also find that 2019 is the most readily alignable year for LLaMa2, while aligning it to 2015 causes the fewest errors measured against all valid answers from 2000 to 2023.

2 TAQA: A QA Dataset for Studying LM Temporal Alignment

To study the temporal chaos of pretrained LMs and how to align them temporally, we curate the TAQA dataset, featuring questions with frequently changing answers in a recent time period (2000-2023). We introduce the formulation of TAQA in §2.1, detail the data construction process in §2.2, and further describe the evaluation metrics in §2.3.

2.1 Problem Formulation

We formally define the task of TAQA as follows: given a question q, which has a set of answers $A = \{(a, t_s, t_e)\}$, where each answer a is only correct when the question is being asked during times $t \in [t^s, t^e]$. Such a dataset can be represented as $\mathcal{D} = \{(q_i, A_i)\}_{i=1}^N$, where at a specific timepoint t, the valid answers of question q shall be $\mathcal{A}(q, t) =$

$$\{a \mid \exists (a, t^s, t^e) \in A \wedge (q, A) \in \mathcal{D}, \text{s.t. } t \in [t^s, t^e]\} \quad (1)$$

We consider a response from a QA system $\hat{a}$ to question q at time t as correct as long as $\hat{a} \in \mathcal{A}(q, t)$. Since only questions with varied answers through history are useful to us, we quantify a question's time-sensitivity as the number of unique answers during a time, i.e.,

$$\mathcal{S}(q, t_s, t_e) = |\{\mathcal{A}(q, t) \mid t \in [t_s, t_e]\}| \quad (2)$$

In practice, we only keep questions such that $\mathcal{S}(q, 2000, 2023) \geq 5\}$ in our TAQA dataset.

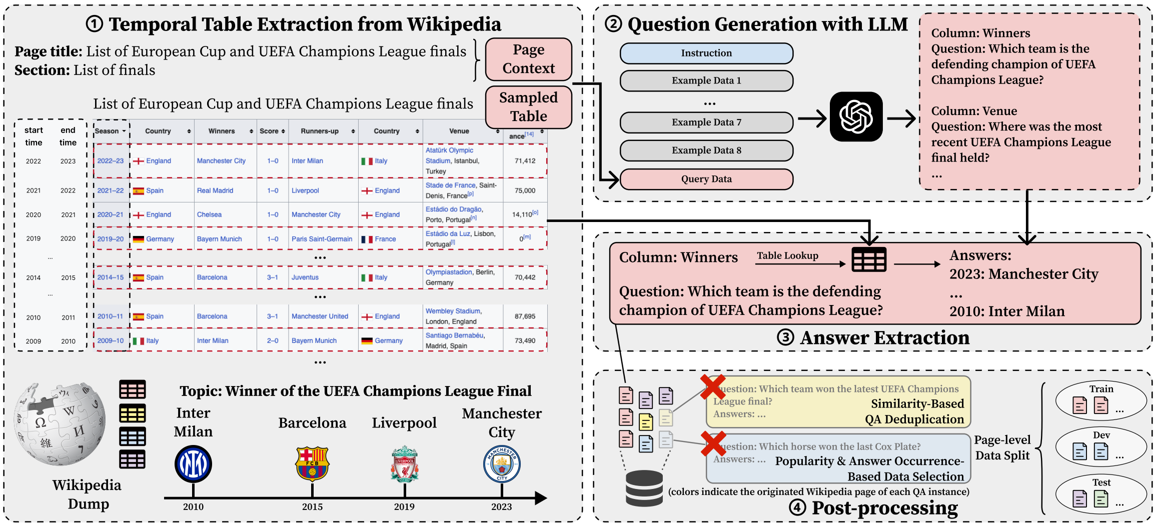

Figure 2: The data construction process of our TAQA dataset.

2.2 Data Construction

To get question q with its multiple temporal answers (a, t_s, t_e), we use tables with temporal columns from the Wikipedia dump to build our TAQA dataset, as the temporal columns can be seen as qualifiers t_s, t_e and other columns can be leveraged to generate QA pairs. The advantage of Wikipedia is that its topics are popular, well-structured, and less biased regarding domains. Furthermore, Wikipedia is used in the pretraining corpus for most modern LMs, ensuring that LMs have learned such knowledge from pretraining. We demonstrate our steps for constructing the TAQA dataset as shown in Figure 2 as follows.

Temporal table extraction from Wikipedia. Given the Wikipedia dump[2], we use the WikiExtractor (Attardi, 2015) tool to extract all tables from it into the CSV format. Then, we use heuristic-based methods to identify the columns denoting temporal information (i.e., (t_s, t_e)) for answers. As we want to gather questions for contemporary knowledge only, we only keep tables with information for every year from 2010 to 2023, resulting in 17,932 tables.

Question generation with LLMs. To generate natural questions with varied styles, we prompt GPT-4 with our manually curated few-shot examples with the tables gathered. We also add each table's Wikipedia page information (title and section names) to the prompt to provide sufficient context. To reduce the query cost, we sample the table and use only the rows corresponding to 2010, 2020, and 2023. To extract answers from the table later, we instruct GPT-4 to generate the column name before each question. After this process, 96,309 question-column pairs are generated in total. Appendix §C presents the detailed prompt.

Answer extraction. With the question-column pairs generated, we use the column names to extract the answers to curate QA pairs from the table directly. When multiple rows have a shared period $[t_s, t_e]$, we treat all answers in those rows as valid. To ensure the question q aligns with our sensitivity requirement $\mathcal{S}(q, 2000, 2023) \geq 5$, we only keep the QA pairs with at least five unique answers.

QA post-processing. To avoid data leakage in our TAQA dataset, we discard QA instances based on their question and answer similarities, measured using BM 25 scores calculated with Lin et al. (2021). When duplicates are found, we keep the instance with fewer words in the question as they are often more natural and popular among real people. Next, we attempt to reduce the dataset's answer bias as we do not want LMs to improve their QA performance by memorizing frequently occurring answers. We begin by identifying questions with answers that are numerical or occur more than 300 times and randomly sample only 10% of them to be kept in our TAQA dataset.

Train-test splitting. Finally, we randomly split the 20,148 questions into train, development, and test sets in TAQA. To better probe LMs' understanding of temporal knowledge like Figure 1, we only put questions that can be answered since 2000 into the dev. and test sets. We also ensure that there is no QA-originated Wikipedia page overlap

[2]We use the English Wikipedia dump created on January 1, 2024.

Split ↓	# Questions	# Tables	# Pages	Popularity	Sensitivity
Train	10,148	6,109	5,751	10912.7	14.6
Dev.	1,000	582	539	5592.2	16.6
Test	9,000	5,290	4,725	6812.5	16.9

Table 1: Statistics of our TAQA dataset. We use Wikipedia pages' averaged monthly pageview to measure questions' popularity, and sensitivity is defined as Equation (2). Popularity and sensitivity are averages across each split.

between the train, dev., and test sets, as we want to avoid LMs improving their QA performance by memorizing the training data's knowledge. The data characteristics are illustrated in Table 1.

2.3 Evaluation Metrics

In TAQA, we focus on the question-answering task and expect models to output short answers. So, following prior QA evaluation (Rajpurkar et al., 2016; Kwiatkowski et al., 2019), we adopt the token-level F1 score to measure the quality of the LM-generated answers given ground truth answers. We further developed three variants tailored for the temporal alignment evaluation.

Target-year F1 Since the ground truth of a question q in our TAQA dataset changes across time, we first calculate the F1 score of an LM-predicted answer $\hat{a}$ to the ground truth in each year between 2000 to 2023, for example, $\mathcal{F}^{2023}(\hat{a}) = \text{F1}(\hat{a}, a_{2023})$ for the LM-predicted answer's accuracy as of in 2023.

Max F1 over history To determine whether an LM can answer questions with historically correct answers if the recent knowledge is unknown to the LM, we also calculate the maximum F1 score among different years' ground truth answers, i.e., $\mathcal{F}_{\text{max}}(\hat{a}) = \max_{2000 \leq i \leq 2023} \mathcal{F}^{i}(\hat{a})$.

Decayed F1 towards target Ideally, temporally aligned LMs' $\mathcal{F}_{\text{max}}$ scores should not be worse than the unaligned one. Furthermore, to differentiate the answers based on recent and outdated knowledge, we softly penalize the outdated answers by calculating the maximum value among F1 scores exponentially decayed by the time gap of the ground truths' year i to the target year j, defined as:

$$\mathcal{F}^{j}_{\text{decay}}(\hat{a}) = \max_{2000 \leq i \leq 2023} (\mathcal{F}^{i}(\hat{a}) \cdot \alpha^{|i-j|}) \quad (3)$$

where $0 < \alpha < 1$ is the decaying factor, which controls acceptance of the outdated answers. When $\alpha = 1$, $\mathcal{F}^{j}_{\text{decay}}$ is the same as $\mathcal{F}_{\text{max}}$, and when $\alpha \to 0$, $\mathcal{F}^{j}_{\text{decay}} \to \mathcal{F}^{j}$. In practice, we set $\alpha = 0.8$ so that a three-year-outdated answer will receive only half of its original F1 score.

3 Aligning LMs towards Recency

In this section, we propose several methods for aligning LMs to recent years. Complementary to existing studies that aim to teach language models up-to-date knowledge (Jin et al., 2022; Meng et al., 2023), our alignment methods focus on how to steer models to reorganize their knowledge and align them to the most-recent possible time.

3.1 Time-Aware Prompting

An intuitive method to make LMs respond based on a specific time is to prompt them explicitly with the time information. We can achieve this by including the target year information. In addition, we also find models can benefit from being prompted with a few demonstration examples that are time-sensitive and with answers in the target time, in a sense that it is inferring the year information from these demonstration answers. We analyze the effects of these elements in Appendix §A Therefore, we propose this time-aware prompting with the year information and a few time-sensitive examples in the prompt to activate models' knowledge of a specific year. Specifically, we append the prompt "as of year y, the answer is" at the end of each question q, where y is the target year we want the LM to be aligned to. For the demonstration examples, we randomly sample 5 instances multiple times from the top-200 popular[3] QA instances in TAQA training set as few-shot examples, and choose the five examples that achieve the best performance on TAQA development set.

The chief advantage of this method is its simplicity, requiring no model updates; it can be used with models that only offer API access. However, prompting does not change models' internal state of time, and for many users, the additional information and few-shot examples are unnatural and add extra costs to inference.

3.2 Target-Year Finetuning

We next explore ways to align LMs temporally by finetuning their parameters. As we want to change

[3]Here, we use the QA's originated Wikipedia page's averaged monthly pageview from 2016 to 2023 to represent popularity.

the models' internal sense of time, we choose not to add any temporal information in the context so that LMs can only generalize when they adjust parameters to model the time information. We refer readers to Appendix §B for the training details.

One critical factor for finetuning is to select the data that can align LMs effectively. In fact, much up-to-date knowledge may not be presented to LMs during pretraining or not be memorized successfully. Tuning LMs with the knowledge they do not know can increase the risk of hallucination (Adlakha et al., 2023; Zhang et al., 2023a), i.e., LMs are encouraged to respond to questions with answers not seen during pretraining[4]. To address this, we propose selecting training data based on the correctness of sampled answers from the LMs. Given a question, we use the time-aware prompting in §3.1 to first sample ten outputs from the to-be-finetuned LM and calculate whether any of the ten samples can hit the ground truth answer. We select the top 5000 examples from the TAQA training set with the largest F1 overlaps between model samples and ground truth answers. By finetuning on question-answer pairs where the model already has some chance of giving the correct answer, we hope to steer it toward the target year rather than "teaching it" entirely new facts. We also explored other data selection methods based on popularity or model confidence, which are reported in §4.3

3.3 Temporal-Adaptive Finetuning

Finetuning to a year assumes the model should answer all questions based on that target time; it, however, fails to take into account a desired graceful degradation: when the LM doesn't have access to the target-year answer, it should revert back to the most recent prior answer.

To achieve this goal, we propose a temporally adaptive finetuning technique, where we dynamically determine the most recent and proper target year for each question. In practice, given a question, we iterate from the pretraining cutoff year (2022 for LLaMa2) to earlier years; for each year, we try using the correctness-based approach described in §3.2 to sample ten outputs, compare with the ground truth answers; if the F1 overlap is larger than a certain threshold (0.7), we will use this year as the target year for this question. As a result, we can adaptively assign each question with its expected target year, and then train the model to output "Based on my latest knowledge for this question from year . . . , the answer is:" before answering the question. Hence, we hope that the LM can learn to pick the target year for each question based on internal knowledge.

[4]This hallucination effect was mentioned in https://news.berkeley.edu/2023/04/24/berkeley-talks-chatgpt-developer-john-schulman and then widely discussed by the community.

4 Experiments

This section describes our experiments on aligning LLaMa2 using our proposed methods on the TAQA dataset. Since the pretraining cutoff of LLaMa2 is September 2022, we try aligning LLaMa2 to both 2022, which it has seen partially, and 2021, which it should have seen entirely. We first introduce several baselines (§4.1), present the effectiveness of aligning LLaMa2 (§4.2), ablate different data selection strategies (§4.3), and finally conduct various analyses to verify the improvement (§4.4).

4.1 Baselines

Unaligned models To demonstrate the performance of LMs after pretraining, we choose several representative LMs, including LLaMa1-65B (Touvron et al., 2023a), LLaMa2-70B (Touvron et al., 2023b), and GPT-3 (Brown et al., 2020) (the "Davinci-002" engine using OpenAI's API). To make them answer questions in the right format, we prompt them with five time-insensitive QA examples, presented in Appendix §A.

Normally aligned models To show the necessity of temporal alignment against other types of alignment in prior work, we additionally test ChatGPT ("gpt-3.5-turbo-0125" engine), which went through an RLHF process (Ouyang et al., 2022), and Tülu2-70B (Ivison et al., 2023), a LLaMa2 model that underwent supervised instruction tuning.

4.2 Results of Recency Alignment

We present the results of aligning LLaMa2 to the years 2021 and 2022 in Table 2, as well as the performance of baseline models. All the **temporally aligned models perform better than unaligned and normally aligned models** in both target years, suggesting the possibility and effectiveness of our proposed temporal alignment. Among all the methods, **finetuning to the target year performs better than just prompting**. Given that we do not include any time information in the prompt for finetuned models, this suggests that the model is learning to

Model ↓	Training Cutoff ↓	$\mathcal{F}^{2022}$	$\mathcal{F}^{2022}_{\text{decay}}$	$\mathcal{F}^{2021}$	$\mathcal{F}^{2021}_{\text{decay}}$	$\mathcal{F}_{\text{max}}$
LLaMA1-65B	Aug. 2022	12.1	26.8	13.2	29.5	54.5
GPT-3	Sep. 2021	10.0	22.9	11.4	25.1	46.5
ChatGPT	Sep. 2021	17.1	32.5	23.8	35.5	46.0
Tülu2-70B	-	16.2	32.2	19.1	35.1	50.1
LLaMA2-70B	Sep. 2022	17.2	34.1	19.8	37.0	55.9
+ prompting to 2021		19.7 (+2.5)	38.9 (+4.8)	25.9 (+6.1)	42.4 (+5.4)	56.7 (+0.8)
+ prompting to 2022		27.4 (+10.2)	40.3 (+6.2)	23.9 (+4.1)	40.5 (+3.5)	54.1 (-1.8)
+ finetuning to 2021		20.8 (+3.6)	39.3 (+5.2)	**29.2 (+9.4)**	**42.5 (+5.5)**	56.7 (+0.8)
+ finetuning to 2022		**27.9 (+10.7)**	**40.7 (+6.6)**	23.7 (+3.9)	40.8 (+3.8)	55.4 (-0.5)
+ adaptive finetuning		25.7 (+8.5)	40.0 (+5.9)	23.5 (+3.7)	40.9 (+3.9)	**57.9 (+2.0)**

Table 2: Performance of temporally unaligned and aligned model on the TAQA dataset. Here, we mainly target 2021 and 2022 for recency alignment, and we evaluate the aligned LMs based on the target-year ($\mathcal{F}^{2022}$ and $\mathcal{F}^{2021}$), temporal-decayed ($\mathcal{F}^{2022}_{\text{decay}}$ and $\mathcal{F}^{2021}_{\text{decay}}$), and historical-max ($\mathcal{F}_{\text{max}}$) F1 scores, as defined in §2.3.

adjust its internal state to activate more of its memorized recent pretraining knowledge. The **benefit of adaptive finetuning is more on historical years**, which is reflected by its best performance on $\mathcal{F}_{\text{max}}$ metric and further confirmed by its performance on a broader range of years shown in Figure 1.

Moreover, comparing the two target years, our models always perform better on 2021. This could be because LLaMa2 did not see all the information in 2022 but we also speculate the information in recent years is less exposed in the pretraining data, as is also shown in Figure 1. We leave the investigation into pretraining data as future work.

4.3 What Data is More Proper for Alignment?

Data selection ↓	$\mathcal{F}^{2022}$	$\mathcal{F}^{2022}_{\text{decay}}$	$\mathcal{F}_{\text{max}}$
All data	19.8	30.9	44.9
Random	19.5 (-0.3)	30.5 (-0.4)	44.3 (-0.6)
Popularity	19.9 (+0.1)	31.3 (+0.4)	45.1 (+0.2)
Model confidence	19.5 (-0.3)	30.7 (-0.2)	44.9 (+0.0)
Model correctness	**20.5 (+0.7)**	**31.8 (+0.9)**	**45.5 (+0.6)**

Table 3: Training data selection analysis with LLaMA2-13B model for optimizing the temporal alignment.

We further analyze the effectiveness of our correctness-based data selection strategy (**Model Correctness**) described in §3.2. We compare it with full-data training, as well as four other selection strategies, including random selection (**Random**), selecting the most popular questions according to page views (**Popularity**), and selecting questions that the model is mostly confident about its answer (**Model Confidence**). All the selection methods select 5000 examples from TAQA training set for fair comparison.

Table 3 compares these methods, proving the superior performance of correctness-based selection. This additionally confirms that the finetuning is activating models' internal knowledge, rather than injecting new knowledge into the model. Otherwise, the popularity-based or full-data training will lead to better performance.

4.4 Improvement Analysis

We provide more detailed analyses of how temporal alignment improves model predictions and confirm that the improvement is not from trivial factors.

Popular knowledge gets aligned better. We first illustrate the association between question popularity and models' improvement on them through temporal alignment, shown in Figure 3 (left). Overall, the target-year F1 scores of both unaligned and aligned models improve as the popularity grows, confirming that popular questions are easier for LMs to answer. Moreover, the improvement from temporal alignment also increases on popular questions, indicating that models have more potential to be aligned on popular topics, since such knowledge is better memorized during pretraining but not triggered due to the temporal chaos.

Improvement is not from memorizing facts in finetuning. The surprising effectiveness of doing temporal alignment through finetuning might be weakened if it is mainly due to some knowledge overlap between the finetuning and testing sets. To confirm this is not the reason, Figure 3 (right) plots the relationship between testing questions'semantic similarity [5] to the training set and models' performance on them. We see that the improvement from alignment does not increase as the similarity grows.

[5]We use sentence-BERT embeddings (Reimers and Gurevych, 2019) to compute the similarity.

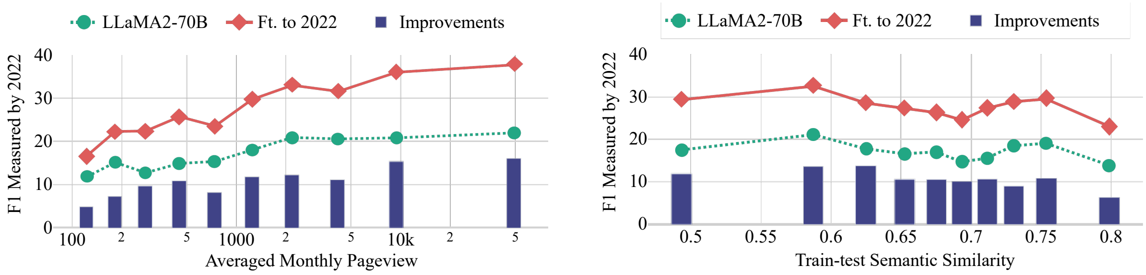

Figure 3: (Left) Relationship between question popularity (measured by the pageviews of their originated Wikipedia page) and models' F1 score on them, as in 2022. (Right) Relationship between testing questions' maximum semantic similarity to the training set and models' F1 score on them, as in 2022.

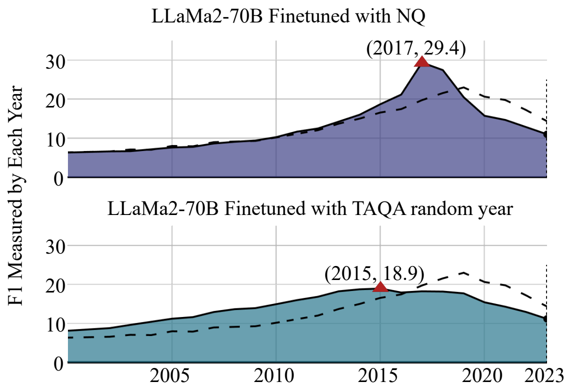

Figure 4: The temporal knowledge distribution of LMs finetuned with the NQ dataset and our TAQA dataset where the answers are randomly sampled.

Method	Metric	7B	13B	70B
Unaligned	$\mathcal{F}^{2022}$	10.4	13.0	17.2
	$\mathcal{F}^{2022}_{\text{decay}}$	21.6	25.6	34.1
	$\mathcal{F}_{\max}$	41.1	45.5	55.9
Finetuned	$\mathcal{F}^{2022}$	17.6 (+7.2)	20.5 (+7.5)	27.9 (+10.7)
	$\mathcal{F}^{2022}_{\text{decay}}$	28.5 (+6.9)	31.8 (+6.2)	40.7 (+6.6)
	$\mathcal{F}_{\max}$	43.1 (+2.0)	46.5 (+1.0)	55.4 (-0.5)
Adaptive	$\mathcal{F}^{2022}$	17.3 (+6.9)	18.9 (+5.9)	25.7 (+8.5)
	$\mathcal{F}^{2022}_{\text{decay}}$	28.8 (+7.2)	32.0 (+6.4)	40.0 (+5.9)
	$\mathcal{F}_{\max}$	44.9 (+3.8)	50.2 (+4.7)	57.9 (+2.0)

Table 4: Effect of LLaMA2 model sizes on temporal alignment, when finetuning to the year 2022.

This implies that memorization is not the reason for better performance on TAQA.

Models are aligned beyond just QA formatting. We further test whether the temporally aligned LMs are learning to activate recent knowledge rather than just learning to answer questions in the correct format. We conduct another two finetunings, with 1) data randomly sampled from NaturalQuestions (NQ, Kwiatkowski et al., 2019) and 2) data randomly sampled from TAQA paired with answers picked from random years. We see from Figure 4 that both of them generally do not have the effect of aligning the model to a recent time. Interestingly, NQ actually aligns the model towards 2017, which we speculate is because the dataset was constructed using much information back then.

4.5 Scaling Trend w.r.t. Model Size

We compare the performance of LLaMa2 70B, 13B, and 7B models before and after finetuning to analyze the effect of model size on temporal alignment. As is shown in Table 4, the improvement of finetuned 7B and 13B models on target-year F1 score ($\mathcal{F}^{2022}$) are both lower than that of the 70B model. Moreover, considering the historical-max F1 score ($\mathcal{F}^{\max}$), the 70B model sees diminishing improvement after temporal alignment finetuning, while 7B and 13B models keep benefiting from finetuning. We speculate that the reason is the 70B pretrained model memorizes more factual knowledge, which might suffer from more forgetting.

5 Further Analysis: To What Year Are LMs Most Readily Alignable?

Aside from aligning LMs to a recent time, we are also curious about which year, historically speaking, an LM is most readily alignable to and achieve the best performance. We investigate this by finetuning LLaMa-70B to different years. Figure 5 shows that 2019 is the best year for aligning the LLaMa2-70B model instead of the year of the pretraining cutoff. Meanwhile, the performance of models finetuned with data whose answers are true in 2010, 2015, and 2019 are all better than in recent years (2022 and 2023). We believe this is because the information about the most recent factual knowledge is too short in existing pretraining corpora, so it is more challenging to align the internal knowledge of LMs to recent times than slightly further to the past.

Furthermore, if considering the historical maximum $\mathcal{F}_{\max}$, the target year of 2015 performs the best among all our target year choices. We argue

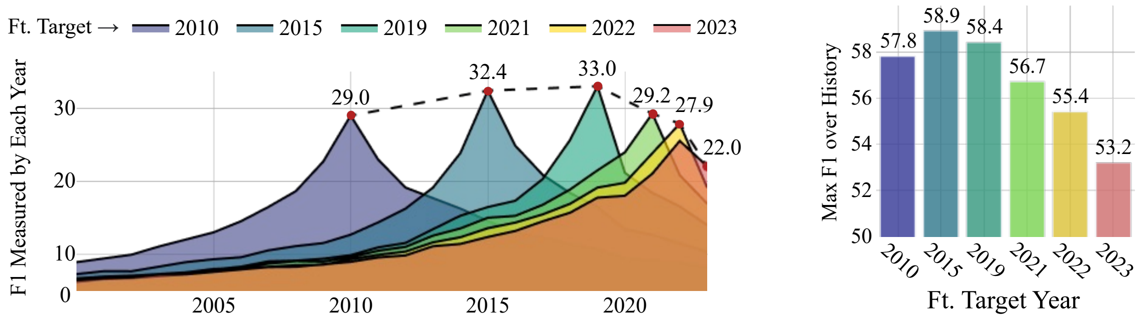

Figure 5: LMs' F1 score in every year between 2000-2023 (left) along with their historical-max F1 scores $\mathcal{F}^{\max}$ (right) after being finetuned to 2010, 2015, 2019, 2021, 2022, and 2023. The dotted line shows the trend of each aligning year's F1 score evaluated with that year's answers.

that the reason could be factual knowledge in 2015 aligns the most with LLaMa2-70B's internal (but unactivated) knowledge learned from pretraining, and therefore, finetuning the LM on TAQA where the answers were true in 2015 causes the least contradiction and thus makes the LM the least likely to produce historically-incorrect answers.

6 Related Work

Temporal Misalignment in LMs Temporal misalignment is a much-studied issue for LMs. For instance, Lazaridou et al. (2021) pointed out that LMs suffer from the temporal gap between their training and deployment data, and Jang et al. (2022) proposed the benchmark of training and evaluating LMs throughout time. Luu et al. (2022) showed that temporal misalignment also hurts LMs' performance on downstream finetuning tasks, and following research indicated this effect can be task-dependent (Agarwal and Nenkova, 2022). Recent research suggested that chaotic pretraining corpus is one of the reasons for LMs' temporal misalignment (Longpre et al., 2023), and LMs can represent temporal knowledge learned from pretraining in their internal states (Gurnee and Tegmark, 2023). These findings open up the possibility of aligning models to a specific time. Different techniques have been proposed to tackle the temporal misalignment issue (Zhang et al., 2023b), including continual pretraining (Jin et al., 2022; Loureiro et al., 2022; Ke et al., 2023), knowledge editing (Mitchell et al., 2022; Meng et al., 2023), and retrieval augmentation (Zhang and Choi, 2023; Vu et al., 2023). These methods edit LMs' knowledge instead of eliciting the knowledge already learned from pretraining, as we aim to do in this work.

Temporal QA and Reasoning Temporal QA and reasoning are tasks where the correct answers change through time, e.g., "Who is the president of the United States?" Prior research tried using knowledge graphs to tackle these tasks (Jia et al., 2021; Saxena et al., 2021; Shang et al., 2022). As modern LMs can memorize factual knowledge, new benchmarks (Chen et al., 2021; Dhingra et al., 2022; Tan et al., 2023; Wei et al., 2023) address the challenge of leveraging LMs to answer questions with contextual temporal mentions provided (e.g., "in 2023", "after" certain events, etc.). However, such question structure can sound unnatural, and these studies do not explicitly focus on questions whose answers have changed in recent years. Other news article-based datasets like RealtimeQA (Kasai et al., 2022) and StreamingQA (Liska et al., 2022), despite having more natural questions, still lack questions whose answers can change multiple times, rather than a single time. In contrast, our TAQA dataset curates natural questions without temporal information but with at least five different answers after 2000. Several approaches have been proposed to tackle the temporal QA and reasoning challenges (Son and Oh, 2023; Tan et al., 2023), yet these approaches require dedicated training for enhancing LMs' understanding of temporal mentions, which are missing in most natural questions.

7 Conclusion

We build TAQA, a QA dataset with time-sensitive questions during a recent time period (2020-2023). Using this dataset, we quantitatively show that LMs can respond to questions quite chaotically in terms of the time it should be based on. We thus propose the concept of temporal alignment and a set of methods for it. Extensive experiments demonstrate the possibility and effectiveness of aligning model internal knowledge temporally, both to a recent time and to a historical time.

8 Limitations

Throughout the development of this project, we identified several limitations that invite additional investigation. Firstly, in our dataset construction, we use GPT-4 to general natural language questions from tabular Wikipedia data without human annotation. While we do adopt quality control policies, it may still be the case that there are noisy generations. These noisy generations, where the question's text content might not be consistent with the quality of a human-written question, may impact the usefulness of such data in the alignment process and thus the effectiveness overall. Secondly, it is difficult to determine the exact date and time that knowledge changes. This, coupled with the fact the exact pretraining corpora of LLaMa2 and similar models are unknown, means we cannot determine whether an LM has observed certain knowledge. This can introduce inaccuracies to scoring when measuring both the original year the model is internally aligned to and the year it is aligned to after our methods are applied. Finally, we note that this study only considers English and English-focused LMs; the applicability of the approaches to other languages is left for future exploration.

9 Ethics Statement

The authors recognize the impact of LMs as components of real-world systems with real-world effects. We wish for our proposed methods make these systems more reliable and useful regarding temporal knowledge for the casual use of LMs and LMs for the professional and research communities. While LMs are still imperfect, our methods of alignment can be used to create more robust systems and mitigate the presentation of inaccurate information to users, which can significantly influence attitudes toward these technologies and the decisions of people using them. We recognize and uphold the ACL Ethics Policy and present the motivation, methodology, and results of our work as accurately as possible.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *arXiv preprint arXiv:2307.16877*.

Oshin Agarwal and Ani Nenkova. 2022. Temporal effects on pre-trained models for language processing tasks. *Transactions of the Association for Computational Linguistics*, 10:904–921.

Giusepppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

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.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

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

Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. 2022. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6237–6250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhen Jia, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Complex temporal question answering on knowledge graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 792–802, New York, NY, USA. Association for Computing Machinery.

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 the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4764–4780, Seattle, United States. 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. 2022. Realtime qa: What's the answer right now? *arXiv preprint arXiv:2207.13332*.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. 2023. Continual pre-training of language models. In *The Eleventh International Conference on Learning Representations*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. 2021. Mind the gap: Assessing temporal generalization in neural language models. In *Advances in Neural Information Processing Systems*, volume 34, pages 29348–29363. Curran Associates, Inc.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. In *Proceedings of the 44th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2021)*, pages 2356–2362.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, D'Autume Cyprien De Masson, Tim Scholtes, Manzil Zaheer, Susannah Young, et al. 2022. Streamingqa: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning*, pages 13604–13622. PMLR.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment.

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022. Fast model editing at scale. In *International Conference on Learning Representations*.

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

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6663–6676, Online. Association for Computational Linguistics.

Chao Shang, Guangtao Wang, Peng Qi, and Jing Huang. 2022. Improving time sensitivity for question answering over temporal knowledge graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8017–8026, Dublin, Ireland. Association for Computational Linguistics.

Jungbin Son and Alice Oh. 2023. Time-aware representation learning for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 70–77, Singapore. Association for Computational Linguistics.

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. *arXiv preprint arXiv:2306.08952*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. 2023. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. *arXiv preprint arXiv:2310.05157*.

Michael J. Q. Zhang and Eunsol Choi. 2023. Mitigating temporal misalignment by discarding outdated facts.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2023a. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Zihan Zhang, Meng Fang, Ling Chen, Mohammad Reza Namazi Rad, and Jun Wang. 2023b. How do large language models capture the ever-changing world knowledge? a review of recent advances. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8289–8311.

Supplemental Material

A Prompting Strategies

We use five strategies to prompt the model and elicit temporal knowledge. In Table 6, we demonstrate each of these strategies with real examples from our dataset that were used to prompt the models used in our experiments.

B Finetuning and Evaluation Details

At first of our finetuning, we did a hyperparameter search using LLaMa2 13 model on the development set of TAQA. We then keep the same set of hyperparameters for all our finetuning experiments. These hyperparameters are shown as follows:

Hyperparameter	Value
Precision	BFloat16
Epochs	2
Learning rate	5e-6
Warmup ratio	0.03
Weigtht decay	Linear
Max. seq. length	128
Batch size	128

Table 5: Finetuing hyperparameters

During the inference of unaligned and normally-aligned LMs, we add five time-sensitive few-shot examples in the prompt as shown in the third row of Table 6 to ensure that the LMs would respond to questions with short answers. When inferring finetuned temporally-aligned models, we do not use specific instructions or prompts but query them with the questions directly.

C Question Generation Details

We use eight few-shot examples to prompt the GPT-4-turbo model (the "gpt-4-1106-preview" engine) for question generation as it is updated with knowledge in 2023, with relatively shorter latency and lower cost. We demonstrate our prompt in Table 7, Table 8, Table 9, and Table 10. We set the temperature to 1.0 for all our queries, hoping to gather more diverse questions in styles and formats. We specifically curated the few-shot examples in vast domains, including sports, films, law, politics, etc. in order to ensure GPT-4 can learn the correct task format in most scenarios when being queried with contents from various Wikipedia pages. We also deliberately add one example (Example 8) without any valid question to be generated to make sure GPT-4 will reject to generate time-insensitive questions when being queried with such unsuitable tables.

D QA Post-Processing Details

After extracting the answers from the Wikipedia tables based on columns selected by GPT-4, we first use a set of heuristic rules to discard noisy QA instances. Specifically, we remove the data that 1) has more than five correct answers per year on average and 2) the averaged answer length is bigger than ten. Therefore, we can ensure that the answers in our TAQA dataset are concise and clear, and short in format.

During QA post-processing, we first use Pyserini (Lin et al., 2021) with Lucene engine to calculate the BM25 scores of each question and answer pairs in the curated dataset. However, since the range of BM25 scores is unbounded, we cannot directly use them to identify duplicate questions with a static threshold. As a result, we use the score normalized with the score between the question/answer to itself to stand for similarity: $\text{Sim}(q_i, q_j) = \text{BM25}(q_i, q_j)/\text{BM25}(q_i, q_i) \in [0, 1]$. A pair of QA instances are marked as duplicates if they fit the following criteria: 1)either question or answer similarity is greater than 0.9, or 2)their question similarity is greater than 0.8 **AND** their answer similarity is greater than 0.5. We set the first criteria's threshold relatively higher because QA instances might have almost the same questions but correspond to totally different answers, e.g., "Who is the President of the United States?" v.s. "Who is the Vice President of the United States".

After de-duplicating the QA data based on the policies above and selecting the instances based on popularity and answer occurrence, as described in §4.3, we further clean the answer texts to confirm there is little noise for LM training and evaluation. In detail, we first discard QA data with answers demonstrating unavailable information, e.g., "N/A", "TBA", "TBD", etc. Afterward, we further clean the redundant HTML tags and Wikitext templates in the answer texts. Finally, as nationalities often appear near peoples' names on Wikipedia tables, we use country codes to identify and clean them as those columns' are designed for questions regarding people but not countries.

Strategy (QA Template)	Template Example
Standard	"prompt": "Answer the following question: What is the role of Karla Estrada in her most recent television series? The answer is:" "completion": " Host"
Time-Aware	"prompt": "Answer the following question: What edition of the Producers Guild of America Awards was last held? As of year 2022, the answer is:" "prediction_text": "33rd"
5-Shot Time-Aware w/ insensitive Examples	"prompt": "Answer the following question: What is the capital of France? As of year 2022, the answer is: Paris Answer the following question: Who wrote Harry Potter? As of year 2022, the answer is: J.K. Rowling Answer the following question: Where did the Titanic sink? As of year 2022, the answer is: Atlantic Ocean Answer the following question: What is the gravity of earth? As of year 2022, the answer is: 9.807 m/s^2 Answer the following question: Is the speed of light faster than the speed of sound? As of year 2022, the answer is: Yes Answer the following question: What edition of the Producers Guild of America Awards was last held? As of year 2022, the answer is:" "prediction_text": "33rd"
5-Shot Time-Aware w/ sensitive Examples	"prompt": "Answer the following question: Which Hindi film has the highest domestic net collection currently? As of year 2022, the answer is: Brahmāstra: Part One – Shiva Answer the following question: Where is the NHL Winter Classic taking place? As of year 2022, the answer is: Target Field Answer the following question: Who are the current drivers for the Mercedes-Benz Formula One team? As of year 2022, the answer is: Lewis Hamilton George Russell Answer the following question: Who received the Player of the Game award for offense in the most recent Rose Bowl Game? As of year 2022, the answer is: Jaxon Smith-Njigba Answer the following question: Where was the final of the last FIFA Club World Cup held? As of year 2022, the answer is: Prince Moulay Abdellah Stadium, Rabat Answer the following question: What edition of the Producers Guild of America Awards was last held? As of year 2022, the answer is:" "prediction_text": "33rd"
Adaptive	"prompt": "Answer the following question: What edition of the Producers Guild of America Awards was last held?" "raw_output": "Based on my latest knowledge for this question from year 2022, the answer is: 33rd"

Table 6: Our prompting strategies and their examples.

Question Generation Template Example

"prompt": "Below is a table in CSV format separated with commas ",". Based on this table's information along with its description and abstracts, please raise up to ten questions that the answers are different in 2010, 2019, and 2023. After raising the question, please get answers for those time points from the table. Please make sure the answers should be totally different without overlapping. Please only raise questions regarding the latest status of those time points instead of the history, while also do not explicitly mention the time information in the question to make them more natural. Do not ask questions cannot be answered based on the information provided in the table. Make sure the full and explicit names of related entities are used in the question based on the description in the prompt and do not use pronouns. Each question should contain only one column's information and no other columns' values shall be mentioned. Please select columns and generation questions based on the Query data, but not the examples. Do not answer the questions in your response. Please reply with the same format as the examples. Do not use a column to generate questions more than once. Do not use other columns' information to be the condition or the clause of the question.
Example 1:
Table description: this table is about List of highest-grossing films Timeline of highest-grossing films.
Table content:
Established,Title,Record-setting gross
1998,Titanic,"$1,843,373,318"
2010,Avatar,"$2,743,577,587"
2010,Avatar,"$2,788,416,135"
2019,Avengers: Endgame,"$2,797,501,328"
2022,Avatar,"$2,923,706,026"

Generated questions for Example 1 asking for information in a specific column:
Column 0: Title
Question 0: What is the highest grossing film of all time?

Column 1: Record-setting gross
Question 1: What is the record-setting gross of the highest grossing film of all time?

Example 2:
Table description: this table is about List of NBA champions Champions.
Table content:
Year,Western champion,Result,Eastern champion,Finals MVP,Coach (1),Coach (2)
2000,"Los Angeles Lakers (1) (25, 12–13)",4–2,"Indiana Pacers (1) (1, 0–1)",Shaquille O'Neal,Phil Jackson,Larry Bird
2010,"Los Angeles Lakers (1) (31, 16–15)",4–3,"Boston Celtics (4) (21, 17–4)",Kobe Bryant,Phil Jackson,Doc Rivers
2019,"Golden State Warriors (1) (11, 6–5)",2–4,"Toronto Raptors (2) (1, 1–0)",Kawhi Leonard,Steve Kerr,Nick Nurse
2023,"Denver Nuggets (1) (1, 1–0)",4–1,"Miami Heat (8) (7, 3–4)",Nikola Jokić,Michael Malone,Erik Spoelstra

Generated questions for Example 2 asking for information in a specific column:
Column 0: Eastern champion
Question 0: Which team was the eastern champion played in the last NBA final?

Column 1: Finals MVP
Question 1: Who's the MVP of the last NBA final?

Example 3:
Table description: this table is about Chris Pratt Filmography, Film.
Table content:
Year,Title,Role,Notes
2000,Cursed Part 3,Devon,Short film
2009,Bride Wars,Fletcher Flemson,
2009,Deep in the Valley,Lester Watts,
2009,Jennifer's Body,Roman Duda,
2009,The Multi-Hyphenate,Chris,Short film
2019,The Lego Movie 2: The Second Part,"Emmet Brickowski, Rex Dangervest",Voice
2019,The Kid,Grant Cutler,
2019,Avengers: Endgame,Peter Quill / Star-Lord,
2023,The Super Mario Bros. Movie,Mario,Voice
2023,Guardians of the Galaxy Vol. 3,Peter Quill / Star-Lord,

Table 7: Our prompting strategies and their examples (part 1).

Question Generation Template Example
Generated questions for Example 3 asking for information in a specific column: Column 0: Title Question 0: What's the movie that Chris Pratt most recently starred in? Column 1: Role Question 1: What's the role that Chris Pratt played in the last movie?   Example 4: Table description: this table is about Judicial Committee of the Privy Council Jurisdiction, Jurisdiction removed. Table content: Country,Date,Abolishing statute,New court of final appeal,Notes The Gambia,1998,1997 Constitution of the Gambia,Supreme Court,"A restructure of the Gambian judiciary by Yahya Jammeh, which made the Supreme Court of The Gambia the highest court instead of being below the Court of Appeal of the Gambia as was the case under the 1970 Constitution of the Gambia." BLZ,2010,"Belize Constitution (Seventh Amendment) Act, 2010",Caribbean Court of Justice, DMA,2015,"Constitution of Dominica (Amendment) Act, 2014",Caribbean Court of Justice, LCA,2023,"Constitution of Saint Lucia (Amendment) Act, 2023",Caribbean Court of Justice, Generated questions for Example 4 asking for information in a specific column: Column 0: Country Question 0: What country abolished the Judicial Committee of the Privy Council of the Uinted Kingdom most recently? Column 1: Abolishing statute Question 1: What is the abolishing statute of the Judicial Committee of the Privy Council of the United Kingdom in the last country that abolished it?   Example 5: Table description: this table is about National People's Congress Membership, Membership of previous National People's Congresses. Table content: Congress,Year,Total deputies,Female deputies,Female %,Minority deputies,Minority % Ninth,1998,2979,650,21.8,428,14.4 Eleventh,2008,2987,637,21.3,411,13.8 Thirteenth,2018,2980,742,24.9,438,14.7 Fourteenth,2023,2977,790,26.5,442,14.8 Generated questions for Example 5 asking for information in a specific column: Column 0: Total deputies Question 0: What is the total number of deputies in the National People's Congress of China? Column 1: Female deputies Question 1: How many female deputies are there in China's National People's Congress? Column 2: Minority % Question 2: What is the ratio of deputy members from minority ethnic groups in China's National People's Congress?   Example 6: Table description: this table is about List of justices of the South Carolina Supreme Court . Table content: Justice,Began active service,Ended active service,Notes Costa M. Pleicones,2000,2017,Became chief justice in 2016 Kaye Gorenflo Hearn,2009,2022,- George C. James,2017,Incumbent,- D. Garrison Hill,2023,Incumbent,- Generated questions for Example 6 asking for information in a specific column: Column 0: Justice Question 0: Who recently became the chief justice of the South Carolina Supreme Court?

Table 8: Our prompting strategies and their examples (part 2).

Question Generation Template Example
Example 7: Table description: this table is about Zlatan Ibrahimović Career statistics, Club. Table content: Club,Season,League (Division),League (Apps),League (Goals),National cup (Apps),National cup (Goals),League cup (Apps),League cup (Goals),Continental (Apps),Continental (Goals),Other (Apps),Other (Goals),Total (Apps),Total (Goals),date_year,start_time,end_time Malmö FF,2000,Superettan,26,12,3,2,—,—,—,—,—,—,29,14,2000.0,2000.0,2000.0 Barcelona,2009–10,La Liga,29,16,2,1,—,—,10,4,4,0,45,21,2009.0,2009.0,2010.0 AC Milan (loan),2010–11,Serie A,29,14,4,3,—,—,8,4,—,—,41,21,2010.0,2010.0,2011.0 LA Galaxy,2019,Major League Soccer,29,30,0,0,2,1,—,—,—,—,31,31,2019.0,2019.0,2019.0 AC Milan,2019–20,Serie A,18,10,2,1,—,—,—,—,—,—,20,11,2019.0,2019.0,2020.0 AC Milan,2021–22,Serie A,23,8,0,0,—,—,4,0,—,—,27,8,2021.0,2021.0,2022.0 Generated questions for Example 7 asking for information in a specific column: Column 0: Club Question 0: Which team does Zlatan Ibrahimović play for? Column 1: League (Division) Question 1: Which league and division does Zlatan Ibrahimović play in? Column 2: League (Goals) Question 2: How many goals does Zlatan Ibrahimović score in league this season? Column 3: League (Apps) Question 3: How many games does Zlatan Ibrahimović play in league this season? Column 4: Total (Goals) Question 4: How many goals does Zlatan Ibrahimović score in total this season?   Example 8: Table description: this table is about List of Denmark women's international footballers List of players. Table content: #,Player,Pos.,Caps,Goals,Debut,Last cap,Medals 94,Signe Højen Andersen,MF,25,0,2000,2004, 23,Heidi Johansen,GK,80,0,2000,2012, 31,Tine Cederkvist,GK,68,0,2000,2011, 24,Gitte Andersen Gitte Andersen (footballer),DF,79,1,2000,2007, 9,Cathrine Paaske Sørensen,MF,121,36,2000,2010, 39,Janne Madsen,MF,61,4,2000,2009, 17,Janni Arnth,DF,93,2,2010,2019,2017 EC 54,Sofie Svava,MF,43,3,2019,2023, 78,Emma Snerle,MF,30,2,2019,2023, 83,Mille Gejl,FW,28,7,2019,2023, 78,Rikke Marie Madsen,FW,30,1,2019,2023, 68,Kathrine Kühl Kathrine Møller Kühl,MF,33,1,2021,2023, Generated questions for Example 8 asking for information in a specific column: No questions can be generated. Query: Table description: this table is about List of European Cup and UEFA Champions League winning managers European Cup and UEFA Champions League winning managers, By year. Table caption: European Cup and UEFA Champions League winning managers* Table content: Final,Nationality,Winning manager,Nation,Club 2010,POR,José Mourinho,ITA,Inter Milan 2019,GER,Jürgen Klopp,ENG,Liverpool 2023,ESP,Pep Guardiola,ENG,Manchester City Generated questions for Query asking for information in a specific column: "

Table 9: Our prompting strategies and their examples (part 3).

Question Generation Template Example
"prediction_text": "Column 0: Final Question 0: In which UEFA Champions League final did the winning manager lead his team to victory? Column 1: Nationality Question 1: What is the nationality of the manager who won the UEFA Champions League? Column 2: Winning manager Question 2: Who is the winning manager of the UEFA Champions League? Column 3: Nation Question 3: From which nation is the winning club of the UEFA Champions League? Column 4: Club Question 4: Which club won the UEFA Champions League?"

Table 10: Our prompting strategies and their examples (part 4).