Can Language Models serve as Temporal Knowledge Bases?

Anonymous EMNLP submission

Abstract

Recent progress in language models as knowledge bases have shown that language models can act as structured knowledge bases for storing relational facts. However, most existing work only considered LM-as-KB paradigm in a static setting, which ignores analysis of temporal dynamics of word knowledge. In this paper, we introduce a new dataset LAMA-TK, aimed at probing language models for temporally-scoped knowledge. We construct cloze statements to query entities and timestamps contained in temporally-scoped facts. To explore the capability of language models as temporal knowledge bases, we propose a temporal scope-aware RoBERTa model and formulate two practical requirements for treating language models as temporal knowledge bases: (i) the ability to store temporal knowledge which contained 1-N relations. (ii) the ability to query stored temporal facts, including implicit temporal facts. Experiments show that conflicting information poses a great challenge to the storage capacity of language models, although language models can memorize millions of temporal knowledge with a relatively high accuracy. Moreover, we show that pre-trained language models can understand implicit temporal knowledge contained in temporal facts and transfer stored knowledge to new queries with similar semantics, even if the query templates are not observed during training.

1 Introduction

Recently, Language models (LMs) such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2020) have been suggested as an alternative to world knowledge bases (Petroni et al., 2019). The parameters of these models appear to store extensive real-world knowledge during training and the stored knowledge can be recalled by filling cloze statements (*e.g.* "Dani Alves plays with [MASK]. –> Barcelona"). As a result, recent work considered language model for tasks such as closed-book question answers (Roberts et al., 2020) , automated fact-checking (Guo et al., 2021) and knowledge-grounded dialogue systems (Liu et al., 2022)

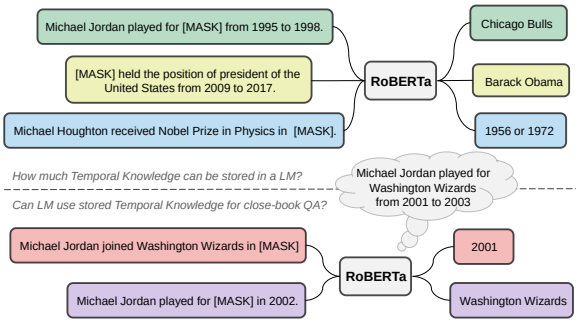

Figure 1: Expansion of LM-as-KB paradigm in temporal domain. We introduce two tasks to further explore the capability of language model. Firstly, we train RoBERTa to memorize millions of temporally-scoped facts and evaluate how much temporal knowledge can be stored into a language model. Secondly, we test the ability of language model to understand implicit temporal knowledge and transfer stored knowledge to new query templates without finetuning.

However, relational facts in world knowledge often change with time. For example, the fact "Giannis Antetokounmpo played for Filathlitikos." is true only from 2011 to 2013. These temporally-scoped facts raise several potential challenges for language model to store temporal knowledge:

Conflicting Information: During training on large textual corpus, the model will inevitably see *1-N* relations, *e.g.*, "Giannis Antetokounmpo played for Filathlitikos / Milwaukee Bucks". By limiting the temporal scopes of facts, the model may see less conflicting information (Dhingra et al., 2022). However, conflicting information still exists, from the players who played for a team in a certain year to the politician who held multiple positions at once. These conflicting facts will hinder the memorizing process and cause the model having low confidences in every correct answers.

Correlation between temporal scopes Temporal facts usually contain temporal scopes (*e.g.*, the start and the end time), and there is a strong correlation between these timestamps. For example, "Shinzō Abe served as the prime minister of Japan from 2006 to 2007." and "Shinzō Abe served as the prime minister of Japan from 2012 to 2014." are two temporally-scoped facts. These facts have the same subject, object and predicate but different temporal scopes. As temporal knowledge bases, LMs need to memorize not only the timestamps associated with the facts, but also the matching relationships between timestamps.

Implicit Temporal Knowledge: Temporally-scoped facts usually contain implicit facts. For example, the fact "Donald Trump served as the president of the United States from 2017 to 2021." contains implicit facts "Donald Trump served as the president of the United States in 2019." and "Donald Trump resigned from president of United States in 2021." These implicit facts are not directly mentioned in temporally-scoped facts, but are implied in them.

Temporal facts are common in real-world knowledge bases like Wikidata. However, existing QA datasets such as LAMA (Petroni et al., 2019), SQuARD (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019) focus on a specific time period, ignoring the temporal dynamics of world knowledge. Some Knowledge Graph Question Answering (KGQA) datasets such as TempQuestions (Jia et al., 2018), CronQuestions (Saxena et al., 2021) contain thousands of temporal questions. But these datasets focus on temporal reasoning and seem too hard for pre-trained LMs without Knowledge Graph Embeddings augmented (Saxena et al., 2021). Moreover, Masked LM evaluation dataset TEMPLAMA (Dhingra et al., 2022) focuses on querying factual object in a single timestamp, ignoring the temporal information contained in real-world facts such as the start time and the end time. Therefore, we propose LAMA-TK (short for **LA**nguage **M**odel **A**nalysis for **T**emporal **K**nowledge), a new dataset for probing LMs for temporal knowledge. LAMA-TK queries temporal knowledge including entity names and specific timestamps (*e.g.* the start time and the end time), and reserves all correct answers for each factual statement. Examples from LAMA, TEMPLAMA and LAMA-TK have been shown in Table 1

Input	Target(s)
LAMA	
Dante was born in [MASK].	Florence
Bailey Peninsula is located in [MASK].	Antarctica
TEMPLAMA	
year: 2012 text: Cristiano Ronaldo plays for _X_.	Real Madrid
year: 2019 text: Cristiano Ronaldo plays for _X_.	Juventus FC
LAMA-TK	
Michael Jordan played for [MASK] from 1995 to 1998.	Chicago Bulls
Michael Jordan played for [MASK] in 2002.	Washington Wizard
Michael Jordan received NBA Most Valuable Player Award in [MASK].	1988, 1991, 1992, 1996, 1998

Table 1: Examples from LAMA, TEMPLAMA and our proposed LAMA-TK. LAMA-TK is a novel dataset of temporal knowledge statements, which takes into account entities, temporal scopes and multiple answers.

In order to comprehensively explore the ability of LMs as temporal knowledge bases, we introduce two fundamental but practical questions for LMs as temporal knowledge bases.

First question: What is the storage capacity of LMs for storing temporal knowledge? What factors will affect the model's storage capacity?

For the first question, we use the LAMA-TK and ask the model to store all temporal entities and temporal scopes contained in temporal facts. Varying from model scale and recording the storage performance of language models. Results show that the storage capacity of language model is directly proportional to the model size, and little affected by pre-training. We also find that storing temporal facts with conflicting information is more challenging than storing static facts or temporal facts without conflicting information.

Second question: Can language model use stored temporal knowledge for closed-book QA? To what extent can LMs understand and use implicit temporal knowledge?

For the second question, we use the LAMA-TK to measure how well can LMs transfer stored temporal knowledge to temporal knowledge queries in zero-shot setting, where the target query templates are not observed during training. These elaborate queries test how well can language model understand and use the stored temporal knowledge, including the ordering and the continuity of temporal scopes. Results show that pre-trained LMs have a fairly good capability to understand implicit temporal knowledge, and can transfer stored temporal knowledge to target queries even if the target query template has never been seen. Moreover, we found that adding an appropriate amount of disturbing to temporal scopes during training can reduce the over dependence on temporal scopes and improve the performance on temporal boundary query.

2 Methods

In this section, we detail the construction of LAMA-TK including the data sources and a set of natural language queries for probing language models as temporal knowledge bases, as well as the models and evaluation metric we use.

2.1 Dataset

LAMA-TK, our new temporally-scoped knowledge probes dataset consists of two parts: a Knolwedge Graph (KG) with temporal annotations and a set of temporal knowledge queries.

2.1.1 Knowledge Sources

CronQuestions CronQuestions (Saxena et al., 2021) is a dataset for Question Answering over Knowledge Graph, including a KG with temporal annotations and a set of temporal questions. There are 323k facts, 125k entities and 203 relations in its KG. We selected top 5 most frequent temporally rich relations and resulted in a KG with 226K facts, 96k entities and 1322 timestamps.

Wikidata Wikidata[1] is a public knowledge base that stored massive structured data. We use the dump of the January 3rd, 2022 version and retrieve facts which have both a start and an end date using SPARQL queries. Following the previous work (Dhingra et al., 2022), we identify the subject and realtion pairs which have multiple objects at different times and select 6 relations with the most such objects. This result in a KG with 497K facts, 260k entities, 1132 timestamps.

2.1.2 Temporal Knowledge Queries

According to the above knowledge sources, we finally construct a KG with 639k facts, 316k entities, 1601 timestamps and 7 relations. Following previous works (Jiang et al., 2020) (Dhingra et al., 2022), we write template for these relations and convert temporal knowledge to natural language statements. For example, the temporal knowledge <Barack_Obama, position_held, president_of_the_United_States, 2009, 2017> was converted into natural language statement "Barack Obama held the position of president of the United States from 2009 to 2017.". Based on these textual statements, we designe targeted cloze-style queries and collect all correct answers for each query. Statistics and example queries for different relations have been shown in Appendix A

[1] www.wikidata.org

Real-world knowledge contains extensive conflicting information, from the players who played for a sport team to the politicians who held multiple positions. Most of previous works do not take into account the negative impact of conflicting information on LMs as knowledge bases. They tend to explore the LM-as-KB paradigm within one-to-one relationships (Heinzerling and Inui, 2021) or only use whether LMs can recall one of the correct answers (*e.g.* Top-K accuracy) to evaluate LMs , without taking into account whether LMs have similar confidences in other correct answers. Therefore, in our proposed LAMA-TK, we additionally mask the subject of each fact to introduce more conflicting information. Among the 2.48M masked factual statements, there are 379K statements with multiple answers.

2.2 Temporal Scope-Aware Language Model

Based on the contextual language model RoBERTa (Liu et al., 2019), we propose a Temporal Scope-aware RoBERTa to explore the capability of language models as temporal knowledge bases.

Prompt-based Temporal Scope Modeling To jointly modeling temporal scopes and text, we manually write *prompt templates* for temporal facts and directly encode temporal scopes in training process. Given a factual sequence of tokens $X = [x_1, x_2, .., x_n]$ and its corresponding temporal scope <ST, ET> (ST i.e. Start Time, ET i.e. End Time). We use prompt template "*from* ST *to* ET" to convert temporal scope to natural language text and incorporate this text into the factual sequence. In this case, the final factual sequence $X' = [x_1, x_2..., x_n, "from", ST, "to", ET]$. See Appendix B for further analysis.

Symbolic Representation However, pre-trained Masked LM can only handle entities whose names are in its vocabulary. This result in its inability to predict entities with multiple words. In this work, we follow (Heinzerling and Inui, 2021) to store entities by symbolic representation, *i.e.*, augmenting the vocabulary of LM and represent all the entities as entries in the vocabulary. The LM will project the final hidden state of the [MASK] token onto the vocabulary and take a softmax over the vocabulary (Heinzerling and Inui, 2021). Although symbolic representation is computationally expensive, it can memorize entities with high accuracy and won't be affected by the length of the entity name.

Memorizing Facts via MLM In this work, we train the model to memorize factual knowledge via Masked Language Modeling (MLM) (Devlin et al., 2019). We use an Entity-level MLM to allow LMs to memorize entities mentioned in factual statements. Formally, Given an input sequence of tokens $X = [x_1, x_2, ..., x_i, x_i + 1, ..., x_n]$ and an two-word entity $e = [x_i, x_i + 1]$. We convert the whole tokens of the entity to one mask token. In this case, the masked sequence of tokens $X' = [x_1, x_2, ..., x_i - 1, [MASK], x_i + 2, ..., x_n]$. Since we use symbolic representation, the masked entity is in the vocabulary of the LM.

2.3 Models

RoBERTa(12L) In this work, we propose a Temporal Scope-aware RoBERTa as the temporal knowledge base. The temporal scope-aware model is initialized from RoBERTa-base (Liu et al., 2019).

RoBERTa(6L) We prepare a 6-layer temporal scope-aware RoBERTa model, initialized from DistilRoBERTa (Sanh et al., 2019), to investigate how knowledge base capability scales with model size.

RoBERTa-randinit(12L) (Heinzerling and Inui, 2021) shows that language models without pre-training can memorize more factual statements than pre-trained models. However, it only focuses on memorizing static and one-to-one relationships. In this work, we also prepare a 12-layer temporal scope-aware RoBERTa without pre-training to further explore the effect of pre-training in a more practical condition.

2.4 Evaluation Metric

As there are many queries with multiple answers, we use the top-K accuracy (Acc@K) to measure how well the model perform on these queries. Top-K accuracy is 1 if any of the top k answers are included in the answer list, and is 0 otherwise. In this work, we use both Acc@1 and Acc@5.

But top-k accuracy is still limited. Acc@K can only measure whether the model can answer the queries correctly, but it cannot indicate how many correct answers the model has memorized (See Appendix C for more details). Therefore, we use Hit at top k (Hit@K) to measure whether the model has memorized all correct answers. For each query, if the masked entity is in the top k answers, Hit@K is 1, otherwise is 0. In this work, we use Hit@5 and Hit@10.

3 Experiments

In this section, we design several experiments to test whether LMs can serve as temporal knowledge bases, including the storage capacity of LMs, as well as the capability of LMs to understand implicit temporal knowledge and use stored temporal knowledge for closed-book QA.

3.1 Storage Capacity

To understand how much temporal knowledge can be stored in a LM, We train prepared models to memorize temporal facts in LAMA-TK. For each fact in LAMA-TK, we mask the subject, object, start time and end time repectively, and generate four masked statements. These masked statements then served as the training data for the LMs to memorize via entity-level MLM, i.e., given the masked statements "[MASK] held the position of president of United States from 2009 to 2017.", the model should predict the masked entity "Barack Obama".

We evaluate the storage capacity of a LM by measuring how much temporal knowledge in training data can be memorized. For example, if the LM's training data contains the temporal fact "Barack Obama held the position of president of the United States from 2009 to 2017.", the model should memorize the fact and recall the correct answer "president of United States." with the query "Barack Obama held the position of [MASK] from 2009 to 2017.".

Note that previous works focus on memorizing static and one-to-one relationships, which makes the task more lightweight, but less practical. In this work, we ask the models to memorize all entities contained in the factual statements. Additionally, masking the subject introduces a large amount of conflicting information, which makes this task more challenging. However, this task is more practical because as a temporal knowledge base, the LM will inevitable see conflicting information, from the players in a sport team to the politician who held multiple positions. Storing conflicting information is a basic function that a knowledge base should have (e.g. storing N-M relations).

Result Red lines in Fig 2 shows the accuracies of statements memorization with different RoBERTa models. Randomly initialized RoBERTa(12L) has the highest recall accuracy for storing temporal knowledge, correctly answer 83 percent of 2.5 million masked statements, while the RoBERTa(6L)

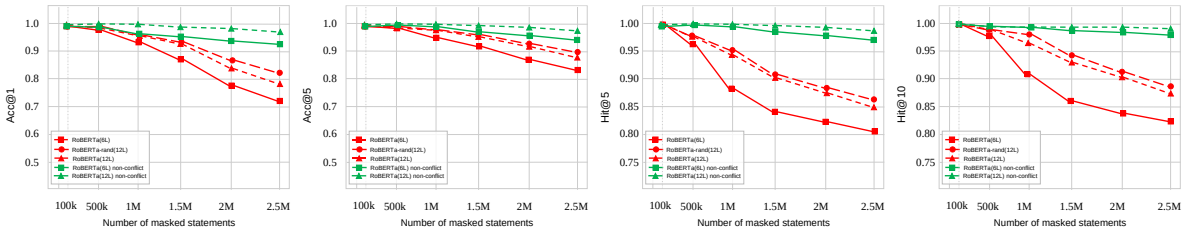

Figure 2: Results of statement memorization. We report Acc@1, Acc@5, Hit@5 and Hit@10 of each model. Green lines show the performances of models trained on LAMA-TK without conflicting information, while red lines show the performances of models trained on LAMA-TK with conflicting information.

has the lowest recall accuracy, with 0.73 Acc@1. As the amount of training data increases, the storage accuracy of all the models gradually decreases. Compared to the RoBERTa(12L), RoBERTa(6L) is more difficult to store 2.5 million masked statements. The result shows that the more parameters the model has, the slower the accuracy of statements memorization decreases. Moreover, by compare with the pre-trained RoBERTa and the randomly initialized RoBERTa, it can be found that randomly initialized LM shows better storage capacity. This is similar to the result of (Heinzerling and Inui, 2021).

Although with the increase of training data, the statements memorization accuracy of all models gradually decrease, the Hit@k of all models remain in a high level. This result shows that LMs can memorize all correct answers despite being affected by the conflicting information.

Influence of Conflicting Information To explore the influence of conflicting information on the storage capacity of language models, we compare models trained on LAMA-TK with and without conflicting information. In LAMA-TK without conflicting information (non-conflict), we remove all masked statements with multiple answers. Then we train RoBERTa(6L) and RoBERTa(12L) on LAMA-TK without conflicting information and record the performance of models.

Green lines in Fig 2 shows the performances of statements memorization without conflicting information. All models can memorize 2 million statements with over 0.95 Acc@1, which is much better than memorizing statements with conflicting information. The drop between memorizing statements with and without conflicting information indicates that the storage capacity of LMs is greatly affected by conflicting information. The accuracy drop of RoBERTa(6L) is more than that of RoBERTa(12L)

Training data	1-1			
	Acc@1	**Acc@5**	**Hit@5**	**Hit@10**
non-conflict	0.9700	0.9910	0.9910	0.9930
conflict	0.8062	0.9366	0.9366	0.9147

Table 2: One-to-one relationship memorization performances of for RoBERTa(12L) trained on 2.48 million masked statements with and without conflicting information.

shows that models with fewer parameters are more susceptible to conflicting information.

Moreover, Table 2 shows the influence of conflicting information on memorizing other one-to-one relationships. The performance drops indicate that conflicting information will hinder the memorizing process of other temporal knowledge, even if it is one-to-one relationship.

3.2 Temporal Boundary Query

From the first experiment, we saw that it is possible for LM to memorize millions of temporal knowledge. We now turn to evaluate the capability of LMs to understand and use temporally-scoped knowledge. First of all we test whether LMs can differentiate between stored timestamps. For example, if the LM has memorized the fact "Barack Obama held the position of president of the United States from 2009 to 2017", the model should recall the start time "2009" with the query "Barack Obama was elected president of the United States in [MASK]" or recall the end time "2017" with the query "Barack Obama resigned from president of the United States in [MASK]".

In order to ensure that the LMs can memorize all required knowledge, we first sample 100k fact statements with the predicate "position held" from LAMA-TK and mask the start time and the end time respectively. This result in 200k masked factual statements. We train RoBERTa models to

Model	Acc@1	Acc@5	Hit@5	Hit@10
RoBERTa(6L)	0.1890*	0.4510*	0.3849*	0.4944*
RoBERTa-rand(12L)	0.1280	0.3260	0.2614	0.3590
RoBERTa(12L)	0.1226	0.3240	0.2689	0.3596
RoBERTa(12L) dynamic mask 10%	0.3774(+0.2658)	0.7042(+0.3802)	0.6628(+0.3939)	0.7740(+0.4144)
RoBERTa(12L) dynamic mask 100%	**0.4879**(+0.3653)	**0.8367**(+0.5127)	**0.7611**(+0.4922)	**0.8838**(+0.5242)

Table 3: Performances of RoBERTa models with and without dynamic time masking on 200k time queries in zero-shot settings. Models above the midrule use orignial masking, while the ones below use dynamic time masking. Green numbers in the brackets show the improvement dynamic time masking brings compared to RoBERTa(12L) with original mask. Highest and second-highest scores among all models are **boldfaced** and underlined. Scores with asterisk are the highest among models with original masking.

memorized all these statements with 0.99 Acc@1.

Next, we write cloze-style templates to query the start time and the end time mentioned in stored facts, such as "S was elected O in [MASK]" and "S resigned from O in [MASK]". We use these queries to test the capability of the model to understand the different between temporal scopes. We conduct this experiment in zero-shot setting , *i.e.*, the target query templates are not observed during training. Zero-shot setting can better show whether the model has knowledge transfer capability and commonsense reasoning ability.

Result The results are shown in the first three rows of Table 3. In the case where the model has fully memorized all required temporal knowledge, the model with fewer parameters performs better. The performance of RoBERTa(12L) is similar to that of RoBERTa-randinit(12L), but both are lower than that of RoBERTa(6L).

Dynamic Time Masking Through the above experiment, we found that the model's capability to query temporal boundary is not satisfactory (low Acc@1). We speculate that this result may be due to the strong correlation between temporal scopes. Original masking makes the model relies too much on the remained timestamp and makes the model difficult to query the masked timestamp without remained timestamp. For example, we use the masked statement "Barack Obama held the position of president of the United State from [MASK] to 2017" to train the model, which make the model's prediction for masked timestamp "2009" excessively relies on the remained timestamp "2017". This makes it hard for LMs to transfer stored temporal knowledge to new queries and result in the model answering these queries with low accuracy.

To verify this conjecture, we inspired by the dynamic masking of RoBERTa (Liu et al., 2019) and design a dynamic time masking. As shown

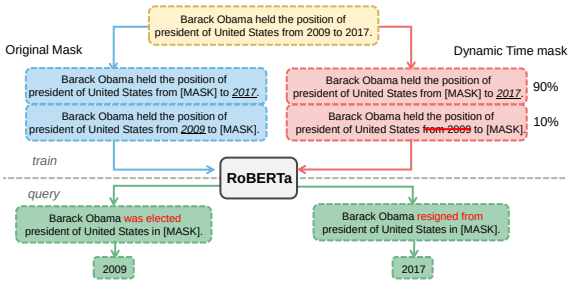

Figure 3: Examples of two types of masking and process of LMs for closed-book QA. The remained timestamps are underlined. The predicates written in red are new query templates, which are not observed during training.

in Figure 3, during constructing masked factual statements, we only mask the specific timestamp *1-k%* of time, and for *k%* of time we mask the specific timestamp and delete the other time information. To avoid using the same time mask in every epoch, we duplicate the training data 10 times so that each statement is masked in 10 different ways over 50 epochs of training. Therefore, each statements was seen with the same mask five times during training.

Dynamic time masking reduces the strong correlation between temporal scopes by adding perturbation to the other temporal information during training. In this experiment, We evaluate RoBERTa(12L) with 10% and 100% dynamic time masking. Table 3. show the performace of these models. By adding 10% perturbation, the accuracy of RoBERTa(12L) significantly increase to 0.3774 Acc@1, 0.7042 Acc@5. Hit@K of RoBERTa(12L) also increase to a high level. We also evaluate RoBERTa with 100% dynamic time masking, which completely ignore the correlation between start time and end time. RoBERTa with 100% dynamic time masking achieved best results both in Acc@k and Hit@k, but 100% dynamic

Model	Parameters	Acc@1 / Acc@5		Hit@5 / Hit@10	
		Template Type		Template Type	
		Original	New	Original	New
RoBERTa(6L)	82M	0.4114 / 0.6521	0.2242 / 0.4115	0.6192 / 0.6993	0.3798 / 0.4540
RoBERTa-rand(12L)	125M	**0.4147 / 0.6868**	0.0131 / 0.0562	**0.6457 / 0.7215**	0.0757 / 0.0518
RoBERTa(12L)	125M	0.3440 / 0.5666	**0.3113 / 0.5020**	0.5281 / 0.6028	**0.4698 / 0.5480**

Table 4: Results on 20k queries with original query templates and new query templates (original query templates: "S held the position of O in T.", new query templates: "S served as O in T."). We report Acc@1/Acc@5 and Hit@5/Hit@10 of each model on two template types.

time masking causes the model unable to associate the start time and end time and unable to handle the facts such as politicians who held one position several times. These results show that dynamic time masking can efficiently help the model reduce the strong correlation between temporal scopes and recall the stored temporal knowledge.

3.3 Implicit Temporal Knowledge Query

In this section, we conduct experiments to test whether LMs can understand the continuity of temporal scopes and use stored implicit temporal knowledge for temporal knowledge queries. For example, if the LM has memorized the fact "Barack Obama held the position of president of United States from 2009 to 2017.", can LM understand that for each year between start time and end time, Barack Obama was the president of United States. Moreover, can LM use this stored implicit temporal knowledge to answer the query "Barack Obama *served as* [MASK] in 2012." even if the template "S served as O in T" is not seen during training.

A controlled experiment is designed for this task. We choose one predicate "position held" and sample all statements generating by template "S held the position of O from ST to ET". To distinguish whether LM answers these queries by using stored knowledge or just by generic association, we inspired by previous work (Heinzerling and Inui, 2021) and add control facts. Given a fact <S, P, O ST, ET>, we add its control <S, P, O', ST', ET'> involves the same subject S and predicate P, but a distinct Object O'. Moreover, we add its control <S, P', O', ST', ET'> involves the same subject S but distinct predicate P' and object O'. For example, control facts for the fact <Barack Obama, Position Held, President of United States, 2009, 2017> are the fact <Barack Obama, Position Held, United States senator, 2007, 2008> and the fact <Barack Obama, award received, Nobel Peace Prize, 2009, 2009>. To correctly answer the query "Barack Obama held the position of [MASK] in 2012.", the model needs to consider both the predicate and the temporal scopes, since there are three distinct objects are associated to "Barack Obama". Every temporal fact has at least one control fact. This process result in 20k factual statements.

Next, We train RoBERTa models to memorize all these fact statements and construct elaborate queries. For each fact, we randomly select one year between the start year and the end year as the timestamp of the query. We do not consider the start year and the end year because these boundary timestamps can bring prompts to the query. Then we use two types of templates to generate queries. Firstly, we use the *Original Template* "S held the position of O in T." to generate queries. This template is also used to generate fact statements for training. Then, we use a *New Template* "S served as O in T" to generate queries. This template has similar semantic information to the original template, but it is not seen during training. We use the *New Template* to test whether LM can transfer stored knowledge into unseen template. This can also be called the robustness of LMs to distinct templates.

Result Table 4 shows the performance of different LMs on two query templates. For *Original Template*, RoBERTa-Randinit(12L) has the highest Acc@k and Hit@k. Compared with RoBERTa(12L), RoBERTa(6L) with fewer parameters performs slightly better. This result is similar to that of previous experiment, which shows that LMs with fewer parameters seem to have a better capability to use stored temporal knowledge.

However, the performance for *New Template* shows a distinct result. In the case the query template is not observed during training, the performance of pre-trained RoBERTa(12L) drops little and remains in a high level, with 0.3113Acc@1 and 0.5480Hit@10. Conversely, the performance of RoBERTa-rand(12L) significantly declines, with only 0.0131Acc@1 and 0.0518Hit@10 .This result shows that pre-trained

LMs have a strong robustness and can transfer stored knowledge to new templates. Compared to RoBERTa(12L), RoBERTa(6L) has lower performace(0.2242 Acc@1 and 0.4540 Hit@10) and drops more(0.4114 Acc@1 to 0.2242Acc@1, 0.6993 Hit@10 to 0.4540 Hit@10). This result shows that the model with more parameters is less affected by new query templates and shows stronger robustness.

4 Related work

Recent research shows that pre-trained langauge model such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), T5 (Raffel et al., 2020), GPT (Radford et al., 2018) can learn extensive world knowledge during pre-training and store these factual knowledge into their parameters. (Petroni et al., 2019) constructs LAMA, a set of cloze-style queries such as "Barack Obama was born in [MASK]. –> Hawaii", to recall factual knowledge contained in Pre-trained LMs such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). Their results show that PLM contains factual knowledge and has strong ability to recall stored knowledge without fine-tuning. (Talmor et al., 2020) proposes eight cloze-stype reasoning tasks such as "Always-Never", "Age COMPARISON" to test different types knowledge in BERT and RoBERTa. While these work focus on probing LM in general domain, (Sung et al., 2021) construct biomedical factual knowledge dataset BioLAMA for probing biomedical LMs, further explore the capability of LM as specific-domain Knowledge Bases. (Heinzerling and Inui, 2021) conduct experiments on RoBERTa to evaluate the ability to store millions of facts involving millions of entities and the ability to query stored facts. Its results provide a proof-of-concept for Langauge Model as Knowledge Bases. Moreover, (Wang et al., 2019) and (Zhou et al., 2020) adopt PLMs on commonsense reasoning tasks, indicating that PLM contains commonsense knowledge. To improve the performance of recalling knowledge, (Petroni et al., 2020) augments PLM with retrived relevant context and improve the performance of cloze-stype question answers. (Jiang et al., 2020) proposes mining-based and paraphrasing-based methods to generate high quality prompts, which significant improve the performance on LAMA.

Within the current paradigm of using Masked Language Models as Knowledge Bases, research has focused more on using Generative Language Models as Knowledge Bases. As Generative Language Models can generate text sequences of any length, they are more convenient as knowledge bases, since they won't be limited by the length of the knowledge. (Roberts et al., 2020) fine-tunes the pre-trained T5 model to three QA datasets WebQuestions (Berant et al., 2013), TriviaQA (Joshi et al., 2017) and NaturalQuestions (Kwiatkowski et al., 2019) without any access to external knowledge to test how much knowledge contained in the LM. The results perform competitively with retrival-based systems and indicates that large pre-trained language models contain vast world knowledge. (Lewis et al., 2021) argues that language models can complete the closed-book QA tasks well is mostly due to the high test-train overlaps. (Wang et al., 2021) designs knowledge memory task and question answering task on low test-train overlaps datasets to evalute the capability of BART (Lewis et al., 2020) serve as knowledge bases for closed-book QA. The results show that closed-book QA is still challenging for BART, both in memorizing the knowledge and answering the questions after memorizing the knowledge. (Dhingra et al., 2022) proposes a time-aware T5 model, which jointly modeling the text with its timestamp, and constrcut a new dataset TEMPLAMA probing LMs for temporal facts. Apart from closed-book QA, (Dai et al., 2022) examine cloze task for BERT to identify the neurons that stored specific fact. The results show the provenance of specific knowledge in parameters of the LM. (Zhu et al., 2020) and (Cao et al., 2021) focus on editing stored knowledge without affecting other unmodified facts. These works further explore the capability of language model and expand the function of language models as knowledge bases.

5 Conclusion

Temporal knowledge is widely exists in real-world knowledge bases. In this work, we extend LM-as-KB paradigm to temporal field and argue that pre-trained LMs have fairly good capability to serve as temporal knowledge bases, in terms of storage capacity, understanding of implicit temporal facts and utilization of stored knowledge. However, our analysis also shows that conflicting information poses great challenges to LM-as-KB paradigm, such as the drop in storage accuracy and the difficulty in recalling multiple answers.

6 Limitations

Our proposed dataset LAMA-TK takes into account temporal scopes of temporal facts and *N-M* relations. But LAMA-TK do not contain questions that require complex temporal reasoning, such as "*First-Last*: [MASK] was the first president of United States.", "*Before-After*: [MASK] was the the president of United States after Barack Obama.". (Saxena et al., 2021) evaluate BERT, RoBERTa, KnowBERT and T5 on CronQuestions which contained 232K such complex questions, but result shows that these large pre-trained langauge models perform very poor (lower than 0.01 Hit@1).

In this work, we propose the temporal scope-aware RoBERTa as the temporal knowledge base. Compared to T5 (737 million parameters), RoBERTa with 12 layers only has 120 million parameters. This makes our experiments lightweight. Moreover, we train RoBERTa to memorize temporal facts via masked language modeling (Devlin et al., 2019). It is possible that incorporating factual knowledge into pre-trained LMs (Sun et al., 2019)(Sun et al., 2020) or augmented LMs with a memory bank (Févry et al., 2020)(Verga et al., 2020) allow language model memorize factual knowledge more efficiently.

Finally, to explore the capability of langauge model to memorize conflicting information (*N-M* relations), we additionally use Hit@K as the evaluation metric to evaluate how many correct answers contained in top-k predictions. However, we do not take into account how to distinguish correct answers from top-k predictions and how many correct answers should be recalled for a query. We plan to investigate these questions in future work.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544. ACL.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6491–6506. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. 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.

Thibault Févry, Livio Baldini Soares, Nicholas FitzGerald, Eunsol Choi, and Tom Kwiatkowski. 2020. Entities as experts: Sparse memory access with entity supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4937–4951. Association for Computational Linguistics.

Zhijiang Guo, Michael Sejr Schlichtkrull, and Andreas Vlachos. 2021. A survey on automated fact-checking. *CoRR*, abs/2108.11896.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1772–1791. Association for Computational Linguistics.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1057–1062. ACM.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Trans. Assoc. Comput. Linguistics*, 8:423–438.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL*

2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, pages 1601–1611. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Patrick S. H. Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1000–1008. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zihan Liu, Mostofa Patwary, Ryan Prenger, Shrimai Prabhumoye, Wei Ping, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Multi-stage prompting for knowledgeable dialogue generation. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1317–1337. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Fabio Petroni, Patrick S. H. Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2020. How context affects language models' factual predictions. In *Conference on Automated Knowledge Base Construction, AKBC 2020, Virtual, June 22-24, 2020*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

Guy D. Rosin, Ido Guy, and Kira Radinsky. 2022. Time masking for temporal language models. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, pages 833–841. ACM.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Apoorv Saxena, Soumen Chakrabarti, and Partha P. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6663–6676. Association for Computational Linguistics.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13,*

2020, pages 3660–3670. International Committee on Computational Linguistics.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Mujeen Sung, Jinhyuk Lee, Sean S. Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4723–4734. Association for Computational Linguistics.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. olmpics - on what language model pre-training captures. *Trans. Assoc. Comput. Linguistics*, 8:743–758.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William W. Cohen. 2020. Facts as experts: Adaptable and interpretable neural memory over symbolic knowledge. *CoRR*, abs/2007.00849.

Cunxiang Wang, Shuailong Liang, Yue Zhang, Xiaonan Li, and Tian Gao. 2019. Does it make sense? and why? A pilot study for sense making and explanation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4020–4026. Association for Computational Linguistics.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. Can generative pre-trained language models serve as knowledge bases for closed-book qa? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3241–3251. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pretrained language models. In *Proceedings of the AAAI Conference of the Artificial Intelligence*, volume 34, pages 9733–9740.

Chen Zhu, Ankit Singh Rawat, Manzil Zaheer, Srinadh Bhojanapalli, Daliang Li, Felix X. Yu, and Sanjiv Kumar. 2020. Modifying memories in transformer models. *CoRR*, abs/2012.00363.

A Statistics and Example Queries

Table 5 shows the statistics and example queries from LAMA-TK. LAMA-TK contains 638,933 temporal knowledge. All these temporal facts are from Wikidata (the Knowledge Graph of CronQuestions(Saxena et al., 2021) is also from Wikidata). For most relations, we use the *prompt template* "*from* ST *to* ET" to convert temporal scopes to natural language texts. However, "award received" is an exception. It is not a durative relation, the start time of the facts is always equal to the end time. Therefore, we use a new *prompt template* "*in* T" to convert these temporal scopes to texts.

B Further Analysis on Prompt-based Temporal Scope Modeling

There are some works focurs on jointly modeling time and text. Time-aware T5(Dhingra et al., 2022) add a time prefix to each text to jointly model time and text. For example, "year:2016 Eden Hazard plays for Chelsea F.C.". TimeBERT(Rosin et al., 2022) adds a time token to the top of the input sequence and design time masking to encode time into the models. For example, "<2022> Joe Biden serves as the Preisdent of the United States of America."

These works focus on modeling text with one timestamp. However, temporal knowledge stored in knowledge bases usually contains temporal scopes (the start time and the end time). Although we can split temporal scopes into years and jointly model the years and texts, this splitting process will lead to a huge increase in factual statements that the model needs to memorize, and introduce a large mount of conflicting information. For example, "Bradley Wiggins played for Ineos Grenadiers in 2010/2011/.../2015.". Section 3.1 has shown that conflicting information can lead to a decrease in the storage capacity of language models. Therefore, we need to find a joint modeling method that can preserve the semantic information of temporal scopes and reducing the introduction of conflicting information.

To this end, we design Prompt-based Temporal Scope Modeling. We use *prompt templates* such as "*from* ST *to* ET" and "*in* T" to jointly model the temporal scopes and factual texts. These prepositions in the prompt templates augment the semantic information of timestamps. Section 3.2 shows that temporal scope-aware RoBERTa preserves the temporal boundary of factual knowledge, and Section 3.3 shows that temporal scope-aware RoBERTa can understand the continuity of temporal scopes without finetuning. These results provide a proof-of-concept that prompt-based template scope modeling can indeed model temporally-scoped knowledge well.

C Limitations of Top-K Accuracy for LM-as-KB tasks

Top-K accuracy indicates that whether the top-k predictions contain correct answers. For example, for the query “Michael Houghton received Nobel Prize in Physics in [MASK].”, we assume that the model recalls one correct answer "1956" at top 1 and recalls another answer "1972" at top 100. Even if the model cannot effectively recall the correct answer "1972", the Acc@1 and Acc@5 to this query is still 1. Therefore, for LM-as-KB tasks, Acc@k can only indicate whether LMs can correctly answer the query, but cannot indicate whether LMs have memorized all correct answers of a query.

In this paper, we use Hit at top k (Hit@K) to evaluate whether LMs have high confidences in all correct answers. For the above example query, the model recalls one correct answer "1956" at top 1 so that Hit@10 for the query "“Michael Houghton received Nobel Prize in Physics in [MASK]. –> *1956*" is 1. However, the model recalls another correct answer "1972" at top 100 so that Hit@10 for the query "Michael Houghton received Nobel Prize in Physics in [MASK]. –> *1972*" is 0. Hit@K provide a more comprehensive result for queries with multiple answers.

D Why not mask the predicate?

In LAMA-TK, we do not mask the predicate because for most temporal facts, there is close association between the predicate and the object. For example, given the object "Nobel Prize in Literature", the model will directly predict the masked relation to be "award received", since the prediction for these relations are hardly affected by entities other than object.

Relation Name	**Template**	**Correct Answers**
	#Relations 7 **#Entities** 316K **#Triples** 638K **#Timestamps** 1601	
educated at	[X] studied at University of Freiburg from 1928 to 1929	Philip Showalter Hench, Bernhard Neumann
position held	Murray Hill held the position of [Y] from 1968 to 1970	Minister for Transport, Minister of Roads,
employer	Emiliano Aguirre worked for University of Granada from [T] to 1974.	1971
member of sport team	Michael Jordan played for Chicago Bulls from 1984 to [T].	1993
award received	Michael Houghton received Nobel Prize in Physics in [T].	1956, 1972

Table 5: Example queries for different relations from LAMA-TK. Different from previous work, we mask not only the object, but also the subject and timestamps. Moreover, we reserve all correct answers for each query. [X], [Y], [T] refers to the masked subject, object, timestamp respectively. The underlined entities are unmasked entities.