PAT-Questions: A Self-Updating Benchmark for Present-Anchored Temporal Question-Answering

Jannat Ara Meem, Muhammad Shihab Rashid, Yue Dong, Vagelis Hristidis
University of California, Riverside
{jmeem001, mrash013, yue.dong}@ucr.edu, vagelis@cs.ucr.edu

Abstract

Existing work on Temporal Question Answering (TQA) has predominantly focused on questions anchored to specific timestamps or events (e.g. 'Who was the US president in 1970?'). Little work has studied questions whose temporal context is relative to the present time (e.g. 'Who was the previous US president?'). We refer to this problem as Present-Anchored Temporal QA (PATQA). PATQA poses unique challenges: (1) large language models (LLMs) may have outdated knowledge, (2) complex temporal relationships (e.g. 'before', 'previous') are hard to reason, (3) multi-hop reasoning may be required, and (4) the gold answers of benchmarks must be continuously updated. To address these challenges, we introduce the PAT-Questions benchmark, which includes single and multi-hop temporal questions. The answers in PAT-Questions can be automatically refreshed by re-running SPARQL queries on a knowledge graph, if available. We evaluate several state-of-the-art LLMs and a SOTA temporal reasoning model (TEMPREASON-T5) on PAT-Questions through direct prompting and retrieval-augmented generation (RAG). The results highlight the limitations of existing solutions in PATQA and motivate the need for new methods to improve PATQA reasoning capabilities.

1 Introduction

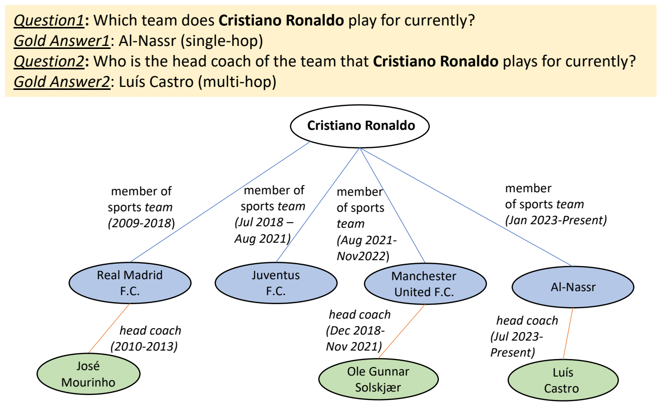
Large language models (LLMs) have demonstrated impressive performance across a wide spectrum of question-answering (QA) domains, thanks to an abundant amount of data spanning different QA tasks such as open-book question answering (OBQA) (Ye et al., 2023; Zhao et al., 2023), knowledge-base question answering (KBQA) (Tan et al., 2023b), and multi-hop reasoning tasks (Kojima et al., 2022; Wang et al., 2023). Their ability to tackle temporal question answering (TQA) has also seen considerable advancements, as evidenced by recent literature (Dhingra et al., 2022; Jia et al., 2018a; Tan et al., 2023a).

However, many past studies on Temporal Question Answering (TQA) focus on questions anchored by specific timestamps or events, such as 'Who was the president of the US in 1985/during World War II?'. In real life, we argue that many of the questions LLMs face are present-anchored without a specific timestamp, representing a crucial yet underexplored category of TQA. We refer to it as Present-Anchored Temporal QA (PATQA), where the time condition is relative to the present time, for example, 'Which team does Cristiano Ronaldo play for currently?'.

PATQA poses challenges due to several factors: (1) LLMs' knowledge becomes **outdated** due to periodic training (He et al., 2022; Zhang and Choi, 2021; Liska et al., 2022). Efforts to mitigate this through retrieval-augmented generation (RAG), providing current documents as context, are also often ineffective (Lewis et al., 2020; Kasai et al., 2022; Vu et al., 2023), as verified by our own experiments with New Bing[1] (using GPT-4, shown in Section 4.2). (2) PATQA can contain **complex temporal relationships** (e.g. before, last, previous) that are challenging. For example, 'Which team did Cristiano Ronaldo play for before the current team?' requires a sequential understanding of temporal expressions 'current' and 'before'. (3) PATQA may require **multi-hop** reasoning that involves temporal reasoning in subsequent hops. For example, tracing Cristiano Ronaldo's current team to its current head coach in 'Who is the head coach of the team that Cristiano Ronaldo plays for currently?' (Cristiano Ronaldo $\rightarrow$ team $\rightarrow$ head coach), requires sequential temporal reasoning followed by multi-hop information integration (head coaches change with time too). *(4)* Creating and **maintaining PATQA benchmarks** is expensive

[1] https://www.bing.com/chat

Findings of the Association for Computational Linguistics ACL 2024, pages 13129–13148
August 11-16, 2024 ©2024 Association for Computational Linguistics



(a) A subgraph from Wikidata around **Subject: Cristiano Ronaldo**

LLM	Release Date	Data Cutoff-date	Answer1	Answer2
Flan-T5-xl	Dec, 2022	Unknown	Real Madrid	José Mourinho
Llama-2-7B	July, 2023	Sept, 2022	Juventus	Juventus (false)
Falcon-7B	May, 2023	June, 2023	Manchester United	Zlatan Ibrahimović (false)
Mistral-7B	Sept, 2023	Unknown	Real Madrid	José Mourinho
GPT-3.5	June, 2020	Jan, 2022	Manchester United	Ole Gunnar Solskjær

(b) LLM responses for the two questions

Figure 1: Illustration of the limitations of the LLMs in answering the present-anchored temporal questions. The LLMs respond with an out-of-date answer (purple) due to knowledge outdating or a false information (red) due to lacking multi-hop PAT reasoning abilities.

because the gold answers to the questions keep changing and manual updates are not sustainable and scalable. Figure 1 illustrates examples that, due to these challenges, current LLMs perform poorly on data in our created PATQA dataset.

We introduce a novel benchmark, referred to as *PAT-Questions*[2], comprising 6172 present time-sensitive factual question-answer pairs that possess the four features we have mentioned above. These challenges require both single and multi-hop temporal reasoning over complex temporal relations to answer correctly. A unique property of PAT-Questions is its capability to automatically update answers over time, resulting in distinct instances for different timestamps. We construct PAT-Questions by leveraging templates derived from time-dependent facts sourced from the Wikidata knowledge base (Vrandečić and Krötzsch, 2014). This allows us to ground our questions on Wikidata facts, thereby ensuring data quality over time by associating a SPARQL query with each question to accurately retrieve answers from the most up-to-date Wikidata.

As far as we know, there are only two datasets which contain present-anchored temporal QA examples, but without complex temporal relations like 'before', 'previous', and have very few multi-hop temporal questions (Kasai et al., 2022; Vu et al., 2023). Further, *these datasets do not offer a way to automatically update the answers over time*, which limits their applicability to future PATQA algorithms.

We benchmark several state-of-the-art (SOTA) LLMs on PAT-Questions, both directly prompting the LLMs with the questions, and in a RAG setting. To retieve documents in RAG, we use Google Custom Search (GCS), following Kasai et al. (2022)'s work, to retrieve relevant documents from up-to-date Wikipedia and Wikidata first, and provide the documents as context along with the initial prompt to the LLMs. We also evaluate the performance of a SOTA temporal reasoning system (Tan et al., 2023a), which fine-tunes the T5-SFT model (Raffel et al., 2020). In their setting, external context in the form of natural language text is provided. In contrast, we consider an open retrieval setting (Nguyen et al., 2016; Rashid et al., 2024b) to retrieve the most relevant context for each question and provide that as context to the LLMs. Our empirical results highlight that the SOTA models significantly struggle on PAT-Questions, especially on multi-hop ones, with EM accuracy ranging from 1.5% to 15.5%.

Our main contributions are:

- We publish a novel PATQA benchmark, PAT-Questions[3], with annotated single-hop and multi-hop questions for two different timestamps (December 2021, December 2023). We provide an automatic answer updating system for the research community to always get up-to-date answers to PAT-Questions.
- We evaluate our benchmark on a wide range of LLMs in direct prompting and RAG settings, and identify limitations of the LLMs in tackling PAT-Questions.

[2]**P**resent-**A**nchored **T**emporal **Questions**

[3]Our dataset and code for self-updates: `https://github.com/jannatmeem95/PAT-Questions.git`

Dataset	Creation	KC	Question Types		PAT	Auto. Ans-update	#ques.
			m-hop	Bef-event reasoning			
Temporal QA & Reasoning Datasets							
TempQuestions (2018a)	Man.-Filt.	Freebase	✓	✓	✗	✗	1271
CRON-QUESTIONS (2021)	Templ.	Wikidata	✓	✗	✗	✗	410k
TimeQA (2021)	Templ.- Wikidata	Wikipedia	✗	✗	✗	✗	20k
SituatedQA-temporal (2021)	Man.-Filt.	Wikipedia	✗	✗	✗	✗	12k
TEMPLAMA (2022)	Templ./ Cloze	Custom-News	✗	✗	✗	✗	50k
StreamingQA (2022)	Man.+Gen	WMT news	✓	✗	✗	✗	410k
TEMPREASON (2023a)	Templ./ Cloze	Wikidata	✗	✓	✗	✗	429k
Present-Anchored Temporal QA Datasets							
REALTIME QA (2022)	News websites	News Articles	✓	✗	some	✗	$\sim$ 5k
FreshQA (2023)	Man.	Google search	✓	✗	377	✗	600
PAT-Questions (ours)	Templ.-Wikidata	Wikipedia	✓	✓	✓	✓	6172

Table 1: Comparison of temporal question-answering datasets. Abbreviations: Man.=created manually, Man.-Filt.=filtered from other datasets, Man.+Gen.=created by crowdsourcing and generated by LLMs, Templ.=created using templates, KC=Knowledge Corpus, PAT=Present Time-Anchored.

- We modify a state-of-the-art temporal reasoning system, Tan et al. (2023a), to answer our PAT-Questions, and experimentally show how it performs on our dataset.

2 Related Work

Temporal Question-Answering Datasets Research on understanding time in texts has led to the development of datasets aimed at enhancing temporal understanding in both knowledge-base question answering (KBQA) and natural language question-answering systems. Prior works on temporal KBQA have led to the creation of datasets like TempQuestions (Jia et al., 2018a), Tequila (Jia et al., 2018b), TimeQuestions (Jia et al., 2021), and CRONQuestions (Saxena et al., 2021), which focuses on integrating temporal data into knowledge bases for ranking entities related to a query (Talukdar et al., 2012; Chang and Manning, 2012). Recent efforts have shifted towards enhancing large language models (LLMs) for time-sensitive reasoning based on natural text only. Datasets like TimeQA (Chen et al., 2021), TEMPLAMA (Dhingra et al., 2022), and TEMPREASON (Tan et al., 2023b) have been introduced to test the ability of LLMs to reason and answer questions that involve understanding explicit temporal context (i.e. 'What team did Cristiano Ronaldo play for in 2021?') or complex temporal relations such as 'before' and 'after' (i.e. 'What team did Cristiano Ronaldo play for before Manchester United?') or to identify time-dependent facts from unstructured text.

Time-sensitive reasoning over Evolving data Existing benchmarks in temporal QA systems focus on static knowledge, annotating questions with single or explicit timestamps, which overlooks the dynamic nature of real-world information where answers can change over time. Notably, SituatedQA-temporal (Zhang and Choi, 2021) and StreamingQA (Liska et al., 2022) have attempted to incorporate temporal context by dating questions and sourcing from recent news, yet they still operate on static snapshots of knowledge. The dynamic REALTIME QA benchmark (Kasai et al., 2022) tests models on current events, however, they exclusively focus on news data and lack emphasis on evolving facts and multi-hop reasoning. FreshQA (Vu et al., 2023) is a contemporary dataset that attempts to update LLMs with current information through time-sensitive questions. Both REALTIME QA and FreshQA rely on the authors to update the answers to reflect new information or changes over time, which limits the datasets' effectiveness in supporting the real-time adaptation of LLMs. In contrast, our dataset, PAT-Questions, can be automatically updated over time ensuring its adaptability and accuracy in real-time, surpassing existing present-anchored datasets. Table 1 shows the comparison among all relevant datasets.

3 PAT-Questions Dataset Construction

We extend the TEMPREASON dataset (Tan et al., 2023a) to construct PAT-Questions. Each question in TEMPREASON follows a time-sensitive template and is annotated with Wikidata IDs for

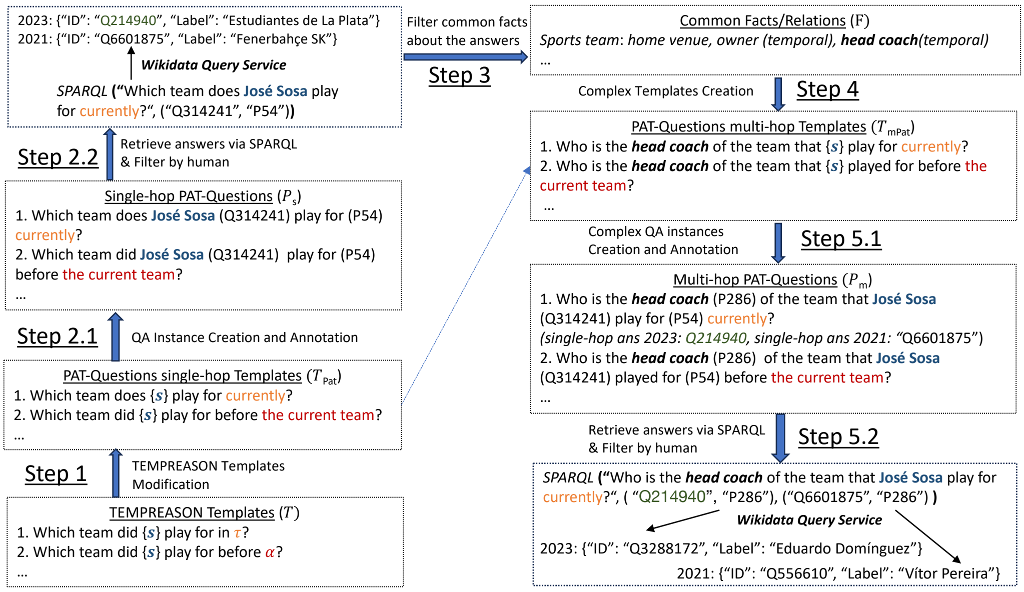


Figure 2: Illustration of PAT-Questions dataset construction following Algorithm 1. Firstly, we modify the time-sensitive templates from the TEMPREASON dataset (Tan et al., 2023a) to build PAT-Questions templates, and following the steps shown in the figure, we create a set of one-hop and multi-hop PAT-Questions with annotated answers for two different timestamps, Dec 2021 and Dec 2023. Here, τ and α refer to a year and an entity respectively.

the primary subject and relation, which facilitates us to generate structured SPARQL queries to automatically update the responses over time. The pre-existing annotations eliminate the need for entity linking as a pre-processing step. We leverage the, the December 31, 2023 Wikidata dump as the knowledge source. We annotate PAT-Questions for two different timestamps of Wikidata to compare the performance of the LLMs. The overall procedure of our data construction is illustrated in Figure 2 and formally defined in Algorithm 1.

Step 1: TEMPREASON Templates Modification TEMPREASON templates by Dhingra et al. (2022), consist of time-sensitive facts $(s, r, o, \tau_s, \tau_e)$, with s representing the subject, r the relation, o the object, and τ_s and τ_e denoting the start and end times of the fact. We adapt two types of TEMPREASON templates (T): i) $(s, r, ?, \tau)$ (where τ lies between τ_s and τ_e) becomes our single-hop PAT template $(s, r, ?, \tau_{cur})$ where τ_{cur} is the current time, and ii) $(s, r, ?, \tau \prec \tau_\alpha)$ (where α is an object related to (s, r) pair facts with distinct τ_s and τ_e, and $\tau \prec \tau_\alpha$ is the time range immediately preceding τ_s) transforms into $(s, r, ?, \tau \prec \tau_{\alpha_{cur}})$ where $\tau \prec \tau_{\alpha_{cur}}$ represents the time range immediately preceding the start time of the current object of the (s, r) pair fact. These rules are outlined formally in Table 2. Our templates (T_{Pat}) are challenging as they don't explicitly specify the

---

Algorithm 1 Construct PAT-Questions Dataset

Require: TEMPREASON dataset, D, TEMPREASON templates T
Ensure: PAT-Questions
1: $T_{Pat} \leftarrow []$ \\ Single-hop templates
2: **for** *each template*, $t \in T$ **do**
3: \\$t = (s, r, \tau)$, or $t = (s, r, \alpha)$
4: **if** $\tau \in t$ **then**
5: $T_{Pat} \leftarrow$ *Replace*(τ, 'currently') $\cup T_{Pat}$
6: **else if** $\alpha \in t$ **then**
7: $T_{Pat} \leftarrow$ *Replace*(α, 'current'+*equiv*(r)) $\cup T_{Pat}$ $\{\, using\ rules\ from\ Table\ 10 \}$
8: S = [subjects(D)] \\ Wikidata subjects for all TEMPREASON questions
9: P_s = *CreateQAInstances*(*SPARQL*(S, T_{Pat})))
10: F = *Filter*(*multiFacts*(P_s))
11: $T_{mPat} \leftarrow []$ \\ Multi-hop templates
12: **for** *each relation* $r_i \in F$ **do**
13: $T_{mPat} \leftarrow$ *Insert*(r_i, T_{Pat}) $\cup$ T_{mPat} $\{\, using\ rules\ from\ Table\ 11 \}$
14: P_m =*CreateQAInstances*(*SPARQL*(S, T_{mPat}))
15: PAT-Questions $= P_s \cup P_m$
16: **return** PAT-Questions

---

current time τ_{cur} or object α, unlike the original TEMPREASON templates. Steps 1-7 of Algorithm 1 depict Step 1, illustrated with examples in Figure 2. Our PAT-Questions single-hop templates are available in Table 10 in Appendix A.

Step 2: Simple QA instances Creation and Annotation We filter the subject entities from original TEMPREASON questions for which the PAT-Questions are valid. Based on Tan et al. (2023a)'s

KB relation, r	**Rule**	**TEMPREASON Template**	**PAT-Questions single-hop Template**
member of sports team (P54)	$(s, r, ?, \tau) \rightarrow (s, r, ?, \tau_{cur})$	Which team did {s} play for in $\boldsymbol{\tau}$?	Which team does {s} play for **currently**?
	$(s, r, ?, \tau \prec \tau_{\alpha}) \rightarrow (s, r, ?, \tau \prec \tau_{\alpha_{cur}})$	Which team did {s} play for before $\boldsymbol{\alpha}$?	Which team did {s} play for before **the current team**?

Table 2: Conversion of the TEMPREASON templates (Step 1) for the 'member of sports team' relation to single-hop PAT-Question templates. TEMPREASON has two templates per relation r and we convert each of the templates following the two rules shown above. For example, s is *Cristiano Ronaldo*, τ is 2021 and α is *Real Madrid F.C.*, the single-hop PAT-Questions become 'Which team does *Cristiano Ronaldo* play for *currently*?' and 'Which team did *Cristiano Ronaldo* play for before *the current team*?'

KB relation, r	**Common relations, r_i**	**Rule**	**PAT-Questions single-hop Template**	**PAT-Questions multi-hop Template**
member of sports team (P54)	home venue (P115), head coach (286)	$(s, r, ?, \tau_{cur}) \rightarrow ((s, r, ?, \tau_{cur}), r_i, ?, \tau_{cur})$	Which team does {s} play for currently (τ_{cur})?	What is the **home venue** of the team that {s} plays for currently? Who is the **head coach** of the team that {s} plays for currently?
		$(s, r, ?, \tau \prec \tau_{\alpha_{cur}}) \rightarrow ((s, r, ?, \tau \prec \tau_{\alpha_{cur}}), r_i, ?, \tau_{cur})$	Which team did {s} play for before the current team ($\tau \prec \tau_{\alpha_{cur}}$)?	What is the **home venue** of the team that {s} played for before the current team? Who is the **head coach** of the team that {s} played for before the current team?

Table 3: Conversion of the PAT-Questions single-hop templates to multi-hop templates (Step 4) for the 'member of sports team' (P54) relation to PAT-Questions multi-hop templates.

approach, we insert the subjects into the single-hop PAT-Questions templates (T_{Pat}) and annotate the questions with the Wikidata IDs of the subjects and relations. Since questions are in natural language, we establish a set of SPARQL query templates to convert each natural language question into its corresponding SPARQL query (see Appendix B). We insert each (s, r) pair into the appropriate SPARQL template, and retrieve the Wikidata ID and NL label of the gold answer using the Wikidata Query Service API(Algorithm 1, lines 8-9). Note that questions are annotated for two different timestamps. We temporally organize the facts linked with (s, r) pairs, fetching the latest objects (o) (current and previous) for the 2023 version, and filtering by end date $\tau_e \leq Dec, 2021$ for the 2021 version.

Step 3: Filter common facts In this step, we randomly select a subset of single-hop question-answer pairs, capturing all templates. We extract all facts (F), including both temporal and non-temporal ones, linked to the answer Wikidata entities (Algorithm 1, line 10). We then filter common facts, i.e., Wikidata triples (s, r, o) shared among these entities, which can be temporal or static. Notably, we prioritize single-hop answer facts over subjects to construct multi-hop templates. This decision is made because single-hop answers span various types, such as sports teams, employers, heads of government/company/organization, etc., resulting in a broader range of facts for our multi-hop questions compared to those associated with the subjects.

Step 4: Complex Templates Creation We generate multi-hop PAT-Question templates (T_{mPat}) by integrating facts from F into the single-hop templates, T_{Pat}, and converting them into natural language following the guidelines outlined in Table 3, where r_i represents one of the relations in F (Algorithm 1, lines 11-13). Note that all answers to the multi-hop templates are grounded on the time the question is posed, denoted as τ_{cur}. The multi-hop PAT templates are listed in Table 11 in Appendix A.

Step 5: Complex QA instances Creation and Annotation For each filtered question in step 2, we insert the subject into its multi-hop template (Algorithm 1, line 14), annotating it with the subject, relation, intermediate entity (gold answer to the single-hop question), and intermediate relation (r_i). Answers are then retrieved following step 3. In this process, we select intermediate entities and relations for insertion into SPARQL templates for

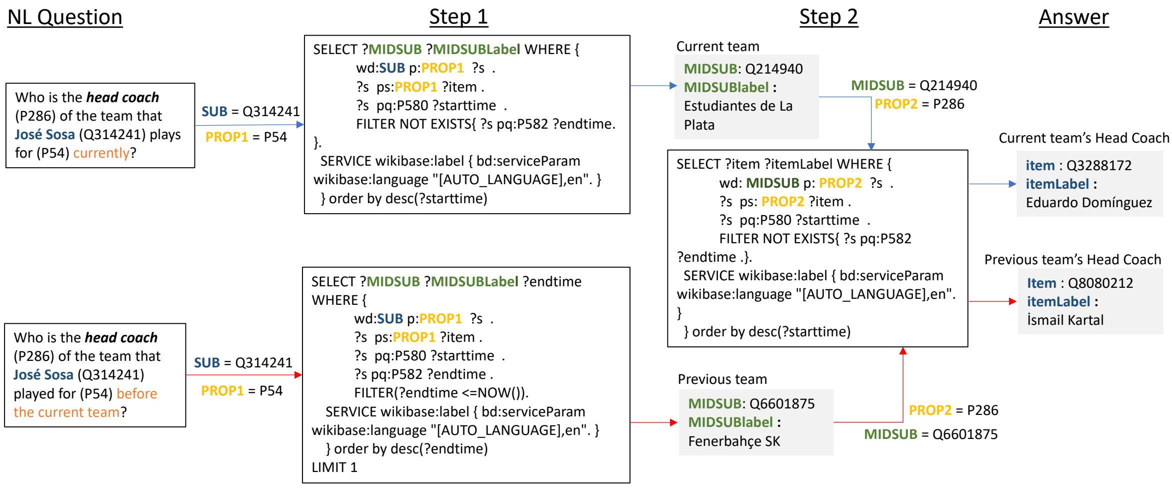


Figure 3: Illustration of automatic answer-updates to two multi-hop PAT-Questions via SPARQL templates

most of the questions (see Appendix C), rather than the original subject and relation. Some questions are filtered out due to missing facts, like spouse (P26), founder (P112), etc. Detailed statistics for **PAT-Questions** are provided in Table 4.

We randomly select 1000 PAT-Questions and manually verify the accuracy of the annotated answers. We also filter out any questions where the answer cannot be retrieved via the SPARQL query.

Type	Category	
	single-hop	multi-hop
current	1442	1617
before-current	1440	1673
Total # questions	2882	3290
	6172	

Table 4: Dataset Statistics of PAT-Questions

Automatically Updating the Answers of PAT-Questions. The questions in our dataset are time-sensitive, with answers expected to change periodically. While the most recent object of Wikidata facts may change, the subject and relation remain constant (Example in Figure 1(a)). Thus, the SPARQL template associated with each question consistently retrieves the latest answer without requiring manual intervention. This functionality empowers users to update the answers to PAT-Questions any time they want. An illustration of the answer update process is provided in Figure 3. Most questions include facts prone to change every six months or longer. To ensure that the research community has the latest answers, we commit to quarterly updates each year, executed through a cronjob running SPARQL queries automatically.

4 Experiments

We conduct experiments on 5 LLMs that have been significantly successful in QA tasks but do not have access to up-to-date world knowledge, including Falcon-7B-Instruct (fal), Flan-T5-XL (Chung et al., 2022), Llama-2-7B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), and GPT-3.5 (Brown et al., 2020) in a direct prompting setting and a RAG setting. We also modify the existing setting of TEMPREASON-T5-SFT by Tan et al. (2023a) to evaluate PAT-Questions in a RAG setting. We compare the results of direct prompting setting at two different timestamps: December 2021 and December 2023. Given that the cutoff date of the LLMs' knowledge is $\geq$ 2021, they should ideally know the answers for December 2021.

4.1 Experimental Setup

Directly Prompting the Pre-trained Models In this experimental setting, we feed each question to the LLMs and instruct the LLMs to answer the question in a few words to avoid verbosity for EM comparisons (see Section 4.1). We use the HuggingFace library for the open-source models and GPT-3.5-Chat API with a temperature of 0. For the 2021 evaluation of the open-source models, we prepend the question with "Assume it is now December 2021," to ensure the fairness of the comparisons with the 2021 gold annotations and with GPT-3.5 for which the cutoff date is January 2022.
Retrieval-Augmented Generation (RAG) In this setting, we augment the LLMs' answer-generation capabilities with retrieval. We retrieve up to five Wikipedia documents for each question us-

	Single-hop				Multi-hop			
	2023		2021		2023		2021	
	EM	F1	EM	F1	EM	F1	EM	F1
Falcon-7B	4.4	5.7	7.8	5.8	2.5	5.6	4.4	6.5
Falcon-7B-w-RAG	8.1	4.9	-	-	4.7	2.9	-	-
Flan-T5-XL	2.0	5.5	2.1	6.0	1.5	5.4	2.8	9.7
Flan-T5-XL-w-RAG	14.9	15.8	-	-	5.1	**9.5**	-	-
Llama-2-7B	8.4	9.0	10.0	11.2	5.3	**8.6**	7.0	9.6
Llama-2-7B-w-RAG	13.9	8.7	-	-	6.6	6.0	-	-
Mistral-7B	7.4	6.4	10.5	7.5	5.7	4.7	6.1	4.8
Mistral-7B-w-RAG	12.7	5.5	-	-	5.9	2.7	-	-
GPT-3.5	**11.7**	11.3	13.6	13.3	**9.3**	7.7	9.7	8.1
GPT-3.5-w-RAG	**15.5**	**16.5**	-	-	**7.6**	6.6	-	-
TEMPREASON-T5-subWiki	12.0	**21.4**	-	-	2.3	7.9	-	-
TEMPREASON-T5-w-RAG	8.3	16.1	-	-	1.5	5.5	-	-

Table 5: The experimental results by EM Accuracy (%) and token-level F1 (%), for two categories of questions of PAT-Questions for two different snapshots of present data (Dec 2023 and Dec 2021)

ing Google Custom Search (GCS) Engine [4], divide each document into chunks of 300 tokens, rank the relevance of these chunks using BM25 and finally assign the top 5 chunks as the retrieved evidence for a question from PAT-Questions dataset. Chunking is necessary in our case because the LLMs that we use have token limitations. We retrieve all the documents for all the questions on the same date (January 16, 2024) to maintain the fairness of our evaluation on the entire dataset. We prompt the LLMs using the question and the retrieved chunks and instruct the LLMs to answer in a few words using the information available in the chunks. We exclusively evaluate this method against December 2023 gold annotations since the retrieved documents contain current information. It would be illogical to retrieve data from a current knowledge source and compare it with outdated gold answers.

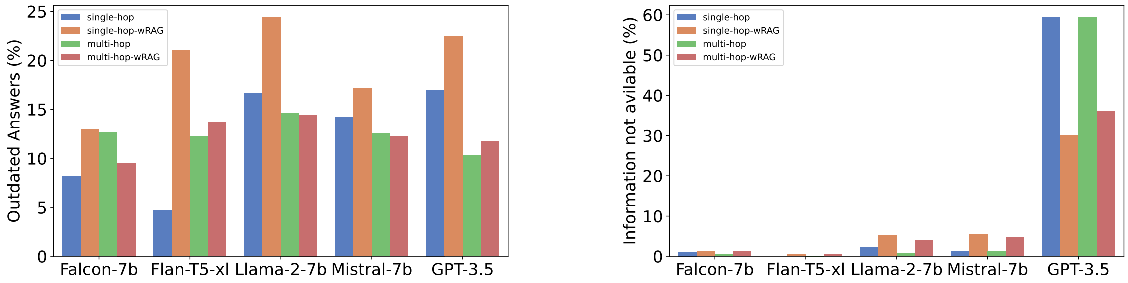
TEMPREASON-T5 Experiments We evaluate our PAT-Questions with Tan et al. (2023b)' T5-SFT model fine-tuned on improving the reasoning capability of the large language model by temporal span extraction. Their Open-book QA (OBQA) setting assumes that the subject entity of the question is already known and they extract the Wikipedia page associated with the subject entity to provide as context to the model. However, this setting is not practical in traditional Open-Retrieval QA settings. As such, we modify their OBQA setting to suit the PATQA problem. We provide the Top 5 BM25 Wikipedia chunks retrieved by GCS for the question as context and, and evaluate their fine-tuned model's performance (TEMPREASON-T5-w-RAG. in Table 5). We also show a comparison with their version of the OBQA setting, meaning we extract the content of the subject entity's current Wikipedia page and provide that as context to the model (TEMPREASON-T5-subWiki in Table 5).

Evaluation Metrics We employ token-level F1 (Rajpurkar et al., 2016) and Chen et al. (2023)'s exact matching (EM) Accuracy metric for the LLMs where if the generated text contains an exact match to the answer or vice-versa, it is considered a correct answer. To address the issue where LLMs might produce an accurate yet differently phrased response to PAT-Questions, such as "Man United" instead of "Manchester United F.C.," resulting in a zero exact match (EM) score, we annotate each answer with all possible aliases from Wikidata using SPARQL queries. For Tan et al. (2023a)'s system, we use traditional Exact Match and F1 to be consistent with their evaluation.

4.2 Results and Discussion

Our findings, presented in Table 5, indicate that pretrained Large Language Models (LLMs) face challenges with PAT-Questions, both single and multi-hop, showing very low EM scores between 1.5% to 15.5% and F1 scores ranging from 2.9% to 16.5%. Accuracy improves with document retrieval, especially for single-hop questions, due to the retrieval of up-to-date and relevant documents. Open-source

[4] https://programmablesearchengine.google.com/



(a) outdated responses (%) by LLMs

(b) 'information not available' responses (%) by LLMs

Figure 4: Error distribution of the incorrect LLM responses

LLMs, significantly underperform in direct prompting settings compared to GPT-3.5 for multi-hop questions. These models considerably benefit from document retrieval due to their lower initial baseline. However, the success of the RAG approach largely depends on the retrieval engine's efficiency, which in our case struggles more with multi-hop than single-hop questions as evidenced by the performance degradation of GPT-3.5 for multi-hop questions. Despite the LLMs' knowledge cut-off date being $\geq$ 2021, the performance compared to 2021 annotations is still very low (though better than the up-to-date annotations). This highlights the LLMs' performance gap in both PAT and multi-hop reasoning. Note that the F1 scores for different models show considerable variation. Flan-T5-XL and GPT-3.5 generally adhere to instructions for concise responses, leading to brief and focused answers. Conversely, other models, including GPT-3.5 in certain instances, tend to produce longer responses, which, despite being accurate, result in lower F1 scores due to their verbosity.

We also compare the performances of TEMPREASON-T5 model with two different contexts: the subject's Wikipedia page and the documents retrieved by GCS. Although the model is specialized in temporal reasoning on the subject's Wikipedia content, it shows low accuracy on both single and multi-hop PAT-Questions. However being fine-tuned on single-hop temporal facts from Wikidata, the model demonstrates comparable results with the open-source LLMs on single-hop questions. The performance degrades significantly for multi-hop questions and open-retrieval RAG settings due to the lack of multi-hop and PAT reasoning capabilities.

We presented a random subset of 50 multi-hop PAT-Questions to New Bing and GPT-4 Web. New Bing accurately answered 9 questions but failed or provided incorrect responses for the remaining 41. GPT-4, on the other hand, correctly answered 6 questions, inaccurately responded to 6, and indicated that information was unavailable for the remaining 38 questions. This comparison highlights the challenges both the services face in handling multi-hop PAT-Questions (see Appendix D).

Error Analysis Figure 4 shows the error distribution of the LLM-generated answers. Figure 4a shows the percentage of outdated answers and Figure 4b shows the percentage of 'information not available' or similar responses out of the incorrect responses of the LLMs based on EM. The responses of Llama-7b, Mistral-7b and GPT-3.5 (especially GPT-3.5) are more grounded to the information available in their parametric memory till the cut-off date for single-hop questions, whereas Flan-T5-XL and Falcon-7b are more likely to generate fake or misinformed responses when not prompted with RAG. Almost all the LLMs struggle with multi-hop reasoning. GPT-3.5 is more cautious in answering present-centric questions and is more likely to respond with 'I do not have real-time information' than responding with an incorrect or outdated answer (see Appendix D and E for more details).

4.3 Additional Experimental Results

Direct Prompting without specifying date We ran experiments on the open-source LLMs with direct prompting without specifying "Assuming it is now December, 2021" in the prompt. The results comparing the generated responses with the December 2021 snapshot are shown in Table 6. The results are lower than that of when the date is specified i.e. the scores shown in Table 5. We also compare the RAG results with 2021 gold annotations and find accuracies going up to some extent (in Table 7). This is due to some questions of our dataset still having the same answers as they were

	Single-hop		Multi-hop	
	EM	F1	EM	F1
Falcon-7B	4.7	6.3	3.6	5.9
Flan-T5-xl	2.1	5.7	2.6	5.7
Llama-2-7B	9.8	10.2	6.0	9.0
Mistral-7B	8.6	7.0	6.1	4.7

Table 6: The experimental results by EM Accuracy (%) and token-level F1 (%), for two categories of questions of PAT-Questions for Dec 2021 snapshot without pretending “Assuming it is December 2021" to the prompt.

	Single-hop		Multi-hop	
	EM	F1	EM	F1
Falcon-7B-w-RAG	8.1	5.1	4.7	2.9
Flan-T5-XL-w-RAG	12.9	14.5	5.7	9.7
Llama-2-7B-w-RAG	13.0	8.7	7.1	6.0
Mistral-7B-w-RAG	12.2	5.0	6.1	2.6
GPT-3.5-w-RAG	14.8	16.3	7.9	6.7
TEMPREASON-T5-subWiki	11.4	20.7	2.4	8.0
TEMPREASON-T5-w-RAG	1.5	5.5	1.5	5.5

Table 7: The experimental results for RAG setting by EM Accuracy (%) and token-level F1 (%), for two categories of questions of PAT-Questions for December 2021 timestamp

in December 2021, however, the LLMs could not correctly answer those. By integrating RAG, those answers have been correctly identified by the LLMs and TEMPREASON-T5.

Advanced Prompting and RAG We conducted additional experiments on multi-hop PAT-Quesions using advanced prompting and RAG techniques. Table 8 shows the experimental results on Chain-of-Thought (CoT) prompting (Wei et al. (2022)), ReAct (Yao et al. (2022)), and Verify-and-edit (Zhao et al., 2023). We observe that CoT results in up to a 6.5% increase in accuracy. This improvement is seen in questions whose answers have remained the same over the last year and for which the LLMs already have the most up-to-date knowledge. However, the overall performance is still very low, clearly indicating that the PAT-Questions dataset poses a significant challenge even for sophisticated prompting methods.

Unlike CoT, ReAct and Verify-and-edit perform worse for some questions, even with external knowledge access, due to temporally misaligned document retrieval. Although the RAG

Method	LLM	EM (%)
Chain-of-Thought (CoT)	Llama-2-7B GPT-3.5	11.2 16.0
ReAct	GPT-3.5	8.8
Verify-and-Edit	GPT-3.5	8.9

Table 8: The EM Accuracy (%), for the multi-hop PAT-Questions (Dec 2023) using advanced CoT prompting (Wei et al. (2022)), and advanced RAG methods (Yao et al. (2022); Zhao et al. (2023)).

	single-hop	multi-hop
Recall@5	19.24%	15.40%
Recall@1	11.1%	8.50%

Table 9: The evaluation of the retrieval quality of Google Custom Search (GCS)

performance improves w.r.t. our GCS retrieval (7.6%), the score is still lower than that of the direct prompting scores for GPT-3.5 (9.3%). These methods show only 35-39% accuracy in multi-hop QA tasks in general, highlighting the gap in multi-hop reasoning. We believe the lower performance with ReAct and Verify-and-edit is due to the additional complexity of present-anchoring. We also observe that these methods often generate reasonable intermediate steps i.e. one-hop questions but the retrieved facts are incorrect, which diverts the next steps from the ideal path (especially for ‘before’/‘previous’ temporal relations).

GCS Retrieval Evaluation Table 9 shows the Recall @top 5 and @top 1 for our Google Custom Search (GCS) retrieval. We observe that the recall scores are very low, especially @top 1, highlighting the gap in traditional retrievers in temporal reasoning, especially in PAT reasoning.

5 Conclusion

In this paper, we introduced a novel self-updating dataset, PAT-Questions, of present-anchored temporal questions requiring both single and multi-hop reasoning on complex temporal relations. We provide a detailed evaluation in both direct prompting and RAG settings of the SOTA LLMs and TEMPREASON-T5 on PAT-Questions, and present the limitations of the LLMs in PATQA. The findings indicate a significant gap in LLMs’ reasoning capabilities when addressing PAT-Questions. We provide an automatic answer updating system for the research community to retrieve the up-to-date answers of PAT-Questions.

Limitations

Our self-updating system depends on an up-to-date knowledge base. We use the Wikidata knowledge base (KB), which may occasionally experience refreshing delays, potentially desynchronizing some gold annotations. Further, we retrieved documents for the PAT-Questions in our RAG pipeline solely using Google Custom Search API. However, this aspect is less significant given that our primary focus is not improving retrieval accuracy. Additionally, the scope of our multi-hop questions is currently limited to 2-hops, which already pose significant challenges for LLMs. We leave 2^+-hop questions for future work.

Ethics Statement

We built our dataset entirely from publicly available information on Wikidata. No personal or restricted data were collected from any source or subject. Although the LLMs may sometimes generate fake information i.e. hallucinate, our experiments do not involve LLMs in creating any harmful content and, thus raise no ethical concern. We adhere to the Code of Ethics with our work.

Acknowledgements

This work was partially supported by NSF grants IIS-2227669 and IIS-1901379.

References

tiiuae/falcon-7b-instruct · hugging face. https://huggingface.co/tiiuae/falcon-7b-instruct. Accessed: 2024-2-13.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Angel X Chang and Christopher D Manning. 2012. Sutime: A library for recognizing and normalizing time expressions. In *Lrec*, volume 12, pages 3735–3740.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation. *arXiv preprint arXiv:2309.01431*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *arXiv preprint arXiv:2108.06314*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. Tempquestions: A benchmark for temporal question answering. In *Companion Proceedings of the The Web Conference 2018*, pages 1057–1062.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018b. Tequila: Temporal question answering over knowledge bases. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1807–1810.

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*, pages 792–802.

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

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*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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

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.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

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. *arXiv preprint arXiv:1606.05250*.

Muhammad Shihab Rashid, Jannat Ara Meem, Yue Dong, and Vagelis Hristidis. 2024a. Ecorank: Budget-constrained text re-ranking using large language models. *arXiv preprint arXiv:2402.10866*.

Muhammad Shihab Rashid, Jannat Ara Meem, and Vagelis Hristidis. 2024b. Normy: Non-uniform history modeling for open retrieval conversational question answering. *arXiv preprint arXiv:2402.04548*.

Apoorv Saxena, Soumen Chakrabarti, and Partha Talukdar. 2021. Question answering over temporal knowledge graphs. *arXiv preprint arXiv:2106.01515*.

Partha Pratim Talukdar, Derry Wijaya, and Tom Mitchell. 2012. Coupled temporal scoping of relational facts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 73–82.

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

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023b. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

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

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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*.

Jinyuan Wang, Junlong Li, and Hai Zhao. 2023. Self-prompted chain-of-thought on large language models for open-domain multi-hop reasoning. *arXiv preprint arXiv:2310.13552*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Hai Ye, Qizhe Xie, and Hwee Tou Ng. 2023. Multi-source test-time adaptation as dueling bandits for extractive question answering. *arXiv preprint arXiv:2306.06779*.

Michael JQ Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. *arXiv preprint arXiv:2109.06157*.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. *arXiv preprint arXiv:2305.03268*.

A Templates details

We show our single-hop and multi-hop templates in Table 10 and 11 respectively. The Wikidata relation distribution over PAT-Questions is shown in Figure 5.

B Dataset Statistics

PAT-Questions contains 2882 single-hop and 3290 multi-hop questions with varied subjects and time-dependent relations from Wikidata. Each question in our dataset has seven common fields associated with it, 'question', 'subject', 'text answers', 'answer annotations', 'relations', 'template', and 'uniq_id'. Multi-hop questions have an extra field named 'intermediate entities' denoting the one-hop answers to the multi-hop questions. An example is shown in Table 12. Figure 5 shows the distribution of different Wikidata relations/properties over our dataset.

KB relation (Wikidata ID)	Single-hop PAT-Questions Templates
member of sports team (P54)	Which team does {s} play for currently? Which team did {s} play for before the current team?
position held (P39)	Which position does {s} hold currently? Which position did {s} hold before the current position?
employer (P108)	Which employer does {s} work for currently? Which employer did {s} work for before the current employer?
political party (P102)	Which political party does {s} belong to currently? Which political party did {s} belong to before the current political party?
head coach (P286)	Who is the head coach of the team {s} currently? Who was the previous head coach of the team {s}?
chairperson (P488)	Who is the chair of {s} currently? Who was the previous chair of {s}?
head of government (P6)	Who is the head of the government of {s} currently? Who was the previous head of the government of {s}?
head of state (P35)	Who is the head of the state of {s} currently? Who was the previous head of the state of {s}?
owner (P127)	Who is the owner of {s} currently? Who was the previous owner of {s}?

Table 10: Single-hop PAT-Questions Templates

C SPARQL Templates

We update the answers to our dataset via running SPARQL queries over Wikidata, and retrieve the most up-to-date answer with respect to knowledge available/updated in Wikidata. Our SPARQL templates are illustrated in Figure 6 and 7. Some cases cannot be handles directly via the SPARQL queries. For example, for some facts 'P1365' qualifier does not exist. For such cases, we have a special clause in our script for extracting the up-to-date answer.

D Sample Prompts and Responses of LLMs

We show the prompt for our experimental settings and some sample responses of the LLMs in Figures 8, 9 and 10. Responses marked in green are the correct answers, blue are the correct answers from a different point in time, and orange are the correct answers to the one-hop questions leading to incorrect or no response to the multi-hop questions. Open-source LLMs tend to provide more relevant and accurate answers compared to when directly prompting the question. This extra instruction helps them ground their answers on the provided time. LLMs' performance improves when one or more relevant passages are provided as context in the RAG setting. As shown in Figures 8 and 9. However, if the retrieved passages are irrelevant, the accuracy may drop, such as for the question "Who was the previous head of the government of India?" in Figure 8, GPT-3.5 fails to answer the question in the RAG setting whereas responds correctly to the same question when asked directly.

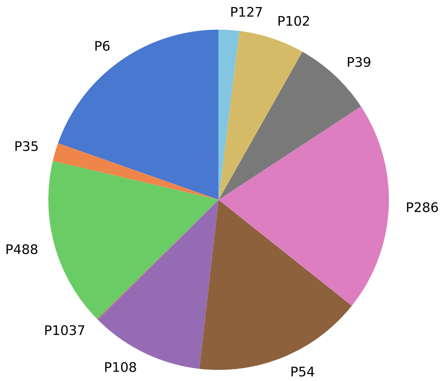
Although utilizing LLMs can be quite expensive, as closed source LLMs charge based on the number of tokens (Rashid et al., 2024a), for our experiments, we tried different prompts such as directly asking the question without any instructions, instructing the LLMs to "Answer the question", "Answer the question in limited words", "Answer the question in a single phrase" etc. Out of the prompts, "Answer the question in limited words" resulted in the best results when compared to the EM accuracy.

Sample Responses of New Bing and GPT-4 Web

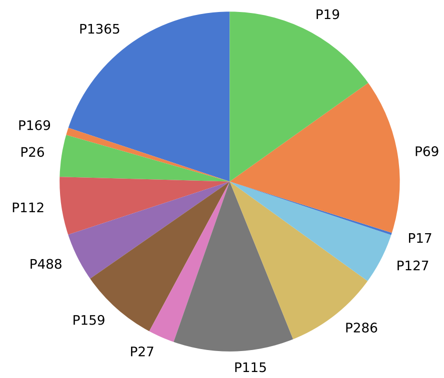
We show some sample responses of BingChat and GPT-4 Web to our multi-hop PAT-Questions in Figure 11. Although these services have access to the current knowledge, they sometimes provide outdated answers (blue), correct one-hop answers but cannot find the two-hop answers (orange), completely false answers (red), or fail to understand the questions and respond with some other facts about the subjects (red).

E Filter for incorrect responses

For our error analysis, we extracted all objects associated with our time-dependent facts along with



(a) Relation distribution over single-hop PAT-Questions



(b) Relation distribution over multi-hop PAT-Questions

Figure 5: Fig. (a) and (b) show the Wikidata relation distributions over PAT-Questions

their aliases to check if an LLM response matches any of the outdated results. To filter responses like 'information not available' or 'cannot respond', we manually went through a subset of LLM responses and picked some keywords that exist in such responses such as 'cannot provide information', 'unknown', 'N/A' etc. Any response that does not fall into the set of extracted outdated objects or 'cannot respond' filter are considered as false or fake, or incomplete response.

Single-hop	
current	SELECT ?item ?itemLabel WHERE { wd:SUB p:PROPERTY ?s . ?s ps:PROPERTY ?item . ?s pq:P580 ?starttime . FILTER NOT EXISTS{ ?s pq:P582 ?endtime .}. SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } } order by desc(?starttime)
	SELECT ?item ?itemLabel WHERE { wd:SUB p:PROPERTY ?s . ?s ps:PROPERTY ?item . ?s pq:P580 ?starttime . ?s pq:P582 ?endtime . FILTER(?endtime>NOW()). SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } } order by desc(?starttime)
previous	SELECT ?item ?itemLabel ?starttime ?endtime WHERE { wd:SUB p:PROPERTY ?s . ?s ps:PROPERTY ?item . ?s pq:P580 ?starttime . ?s pq:P582 ?endtime . FILTER(?endtime <=NOW()). SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } } order by desc(?endtime) Pick the one (or multiple) item (s) with the most recent endtime.

Figure 6: SPARQL templates for single-hop PAT-Questions' answer-update

KB relation (Wikidata ID)	Multi-hop PAT-Questions Templates
head coach (P286)	Who is the head coach of the team that {s} plays for currently? Who is the head coach of the team that {s} played for before the current team?
home venue (P115)	What is the home venue of the team that {s} plays for currently? What is the home venue of the team that {s} played for before the current team?
owned by (P127)	Who is the owner of the team that subject play for currently? Who is the owner of the team that {s} played for before the current team?
replaces (P1365)	Who was the last person to hold the position that {s} holds currently? Who was the last person to hold the position that {s} held previously?
spouse (P26)	Who is the spouse of the current chair/owner/head of the state/head of the government of {s}? Who is the spouse of the previous chair/owner/head of the state/head of the government of {s}?
educated at (P69)	Which school did the current head coach/chair/owner/head of the government/head of the state of {s} attended? Which school did the previous head coach/chair/owner/head of the government/head of the state of {s} attended?
headquarters (P159)	Where is the headquarters of the team that {s} plays for currently? Where is the headquarters of the political party {s} belongs to currently? Where is the headquarters of the current/previous owner of {s} located at? Where is the headquarters of the current/previous employer of {s}? Where is the headquarters of the team that {s} played for before the current team? Where is the headquarters of the political party that {s} belonged to before the current political party?
chair (P488)	Who is the chair of the current employer of {s}? Who is the chair of the political party which the {s} belongs to currently? Who is the chair of the previous employer of {s}? Who is the chair of the political party which the {s} belonged to before the current political party?
chief executive officer (P169)	Who is the chief executive officer of the current employer of {s}? Who is the chair of the previous chief executive officer of {s}?
country of citizenship (P27)	What is the country of citizenship of the current chair/head coach of {s}? What is the country of citizenship of the previous chair/head coach of {s}?
country (P17)	Which country is the current owner of {s} from? Which country is the previous owner of {s} from?
birthplace (P19)	Where was the current head coach/head of the government/chair/owner of {s} born? Where was the previous head coach/head of the government/chair/owner of {s} born?

Table 11: Multi-hop PAT-Questions Templates

‘question’: “Who is the spouse of the current head of the state of Indonesia?”,
“subject”: {“subject”: "Q252", “subjLabel": "Indonesia"},
"text answers": [“Iriana"],
"answer annotations": [{“ID": "Q17410605", “Label": "Iriana"}],
"intermediate entities": [{“ID": "Q3318231", “Label": "Joko Widodo" }],
“relations": [“P35", “P26"],
“template": “Who is the spouse of the current head of the state of {subject}?”,
“uniq_id": 2986

Table 12: An example multi-hop question from our PAT-Question dataset

Multi-hop	
current (all properties except P1365)	SELECT ?item ?itemLabel WHERE { wd:MIDSUB p:PROPERTY ?s . ?s ps:PROPERTY ?item . ?s pq:P580 ?starttime . FILTER NOT EXISTS{ ?s pq:P582 ?endtime .}. SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } } order by desc(?starttime)
	SELECT ?item ?itemLabel WHERE { wd:MIDSUB p:PROPERTY ?s . ?s ps:PROPERTY ?item . ?s pq:P580 ?starttime . ?s pq:P582 ?endtime . FILTER(?endtime>NOW()). SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } } order by desc(?starttime)
previous (all properties except P1365)	SELECT ?item ?itemLabel ?starttime ?endtime WHERE { wd:MIDSUB p:PROPERTY ?s . ?s ps:PROPERTY ?item . ?s pq:P580 ?starttime . ?s pq:P582 ?endtime . FILTER(?endtime <=NOW()). SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". } } order by desc(?endtime) Pick the one (or multiple) item (s) with the most recent endtime.

current & previous
(P1365)
	SELECT ?holder ?holderLabel ?item ?itemLabel WHERE {
	wd:SUB p:PROPERTY ?s .
	?s ps:PROPERTY ?holder.
	?s pq:P580 ?starttime .
	OPTIONAL {
	?s pq:P1365 ?item.
	}
	SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
	}order by desc(?starttime)
	Pick the first item for which holder == MIDSUB

Figure 7: SPARQL templates for multi-hop PAT-Questions’ answer-update

Prompt	Question	Falcon-7B	Flan-T5-XL	Llama-2-7B	Mistral-7B	GPT-3.5	WikiData Gold Answer
Answer the question in limited words.	Who was the previous head of the government of India?	Jawaharlal Nehru	Narendra modi	Manmohan Singh	The previous head of the government of India was Dr. Manmohan Singh.	Manmohan Singh	Manmohan Singh
	Who is the head of the government of Catalonia currently?	The current head of the government of Catalonia is Quim Torra	Josep Maria Bartomeu	Quim Torra	The head of the government of Catalonia currently is Quim Torra	Pere Aragonès	Pere Aragonès i Garcia
	Which team does Volodymyr Gerun play for play for currently?	Volodymyr Gerun currently plays for Dynamo Kyiv	FC Dynamo Kyiv	Volodymyr Gerun currently plays for Shakhtar Donetsk	Volodymyr Gerun plays for FC Dynam	Volodymyr Gerun currently plays for Sendai 89ers	Sendai 89ers
	Which team did Josè Ernesto Sosa play for before the current team?	Sosa played for the New York Yankees	Los Angeles Dodgers	Before joining his current team, Josè Ernesto Sosa played for the team Independiente	Josè Ernesto Sosa played for the team Boca Juniors before the current	Trabzonspor	Fenerbahçe SK
	Which employer did Nada Al-Nashif work for before the current employer?	Nada Al-Nashif worked for Google before her current employer	Emirates	Nada Al-Nashif worked for the World Bank before her current employer	Nada Al-Nashif worked for the United Nations High Commissioner for Refuge	United Nations	UNESCO
Answer the question in limited words using information with the help of the external documents followed by the user question. \n Question: … \n Documents: …	Who was the previous head of the government of India?	The previous head of the government of India was Jawaharlal Nehru	Sardar Vallabhbhai Patel	The previous head of the government of India was Narendra Modi	The previous head of the government of India was Dr. Manmohan Singh.	The previous head of the government of India is not mentioned in the provided documents.	Manmohan Singh
	Who is the head of the government of Catalonia currently?	Pere Aragones of the Republican Left of Catalonia, following his election in 21 May 2021 after the 2021 Catalan elections.	Pere Aragones	The current head of the government of Catalonia is Pere Aragones of the Republican Left of Catalonia.	The current president of the government of Catalonia is Pere Aragones of the Republican Left of Catalonia, following his election in 21 May 2021 after the 2021 Catalan elections.	The current head of the government of Catalonia is Pere Aragones.	Pere Aragonès i Garcia
	Which team does Volodymyr Gerun play for play for currently?	Volodymyr Gerun is a professional basketball player for Sendai 89ers of the B.League	Sendai 89ers	Volodymyr Gerun currently plays for Sendai 89ers of the B.League	Volodymyr Gerun plays for Sendai 89ers of the B.League.	Volodymyr Gerun currently plays for Sendai 89ers	Sendai 89ers
	Which team did Josè Ernesto Sosa play for before the current team?	Estudiantes de La Plata	Estudiantes de La Plata	Josè Ernesto Sosa played for Estudiantes de La Plata before current team.	Josè Ernesto Sosa played for the team Estudiantes de La Plata	Estudiantes de La Plata	Fenerbahçe SK
	Which employer did Nada Al-Nashif work for before the current employer?	Nada Al-Nashif worked for the United Nations International Labour Organization	International Labour Office	Nada Al-Nashif worked for the International Labour Office before her current employer.	Nada Al-Nashif worked for the United Nations High Commissioner for Refuge	The documents provided do not mention the previous employer of Nada Al-Nashif.	UNESCO

Figure 8: Responses of LLMs to single-hop PAT-Questions when asked without and with RAG.

Prompt	Question	Falcon-7B	Flan-T5-XL	Llama-2-7B	Mistral-7B	GPT-3.5	WikiData Gold Answer
Answer the question in limited words.	Who is the head coach of the football club that José Sosa play for currently?	New York Yankees	José Luis ngel de la Rosa	Diego Armando Maradona	José Ernesto Sosa is a professional footballer who plays as a midfielder.	I cannot provide you with the current head coach of José Ernesto Sosa's team.	Eduardo Domínguez
	Who is the head coach of the team that Oleksandr Filippov played for before the current team?	Oleksandr Filippov plays for Dynamo Kyiv	Oleksandr Yakovlev	The head coach of Oleksandr Filippov's current team is Andriy Shevchenko	Oleksandr Filippov is a Ukrainian professional footballer who plays as	It is not known which team Oleksandr Filippov currently plays for.	Bernd Hollerbach
	Who is the chief executive officer of the current employer of Ken Murphy?	Ken Murphy	John Sutter	Satya Nadella	The chief executive officer of the current employer of Ken Murph is Mr. John	Ken Murphy	Ken Murphy
	Who is the spouse of the current head of the state of Indonesia?	The spouse of the current head of the state of Indonesia is Tante Yusuf	Susilo Bambang Yudhoyono	Iriana Joko Widodo	The current head of the state of Indonesia is Joko Widodo.	I am sorry, but as an AI language model, I don't have access to real-time information.	Iriana
	Who is the chair of the previous employer of Nada Al-Nashif?	Nada Al-Nashif	Sheikha Sabah Al-Ahmed Al-Sabah Al-Sa	John Smith	Nada Al-Nashif is the chair of the previous employer of Nada Al.	There is no information provided about the previous employer of Nada Al-Nashif	Audrey Azoulay
Answer the question in limited words using information with the help of the external documents followed by the user question. \n Question: ... \n Documents:	Who is the head coach of the football club that José Sosa play for currently?	José Ernesto Sosa plays for the New York Yankees.	José Luis ngel de la Rosa	The head coach of the team that Jos\u00e9 Ernesto Sosa plays for is Diego Armando Maradona.	José Ernesto Sosa is a professional footballer who plays as a midf	The head coach is not mentioned in the provided documents.	Eduardo Domínguez
	Who is the head coach of the team that Oleksandr Filippov played for before the current team?	Oleksandr Filippov is currently playing for the Ukrainian Premier League club Desna Chernihiv	Oleksandr Ryabokon	The head coach of the team that Oleksandr Filippov play for is not mentioned in the provided documents.	Oleksandr Ryabokon is the head coach of the team that Oleksandr Filippov play for play for currently.	The head coach of the team Oleksandr Filippov currently plays for is not mentioned in the provided documents.	Bernd Hollerbach
	Who is the chief executive officer of the current employer of Ken Murph?	Ken Murph	John Sutter	Ken Murph is the chief executive officer of Tesco	The chief executive officer of the current employer of Ken Murph is Mr. John	Ken Murph is the CEO of Tesco.	Ken Murph
	Who is the spouse of the current head of the state of Indonesia?	The spouse of the current head of the state of Indonesia is Joko Widodo	Joko Widodo	Iriana Widodo	The current head of the state of Indonesia is Joko Widodo. His spouse is Iriana	Iriana	Iriana
	Who is the chair of the previous employer of Nada Al-Nashif?	Roman Badanin	Vladimir Ivanovich Chernomyrdin	Igor Kucher	The chair of the previous employer of Roman Badanin is Mr. K.	Vladimir Kovalchuk	Audrey Azoulay

Figure 9: Responses of LLMs to multi-hop PAT-Questions when asked without and with RAG.

Prompt	Question	Falcon-7B	Flan-T5-XL	Llama-2-7B	Mistral-7B	WikiData Gold Answer
Assuming it is now December 2021, answer the question in limited words.	Who was the previous head of the government of India?	Manmohan Singh	Narendra Modi	Narendra Modi	The previous head of the government of India was Narendra Modi	Manmohan Singh
	Who is the head of the government of Catalonia currently?	Pere Aragonès	ida a ferenc igual	Quim Torra	Pere Aragonès	Pere Aragonès i Garcia
	Which team does Volodymyr Gerun play for play for currently?	Volodymyr Gerun currently plays for Dynamo Kyiv	FC Dynamo Kyiv	Volodymyr Gerun currently plays for Shakhtar Donetsk	Volodymyr Gerun is a Ukrainian footballer who plays for FC Dynam	Sendai 89ers
	Which team did Josè Ernesto Sosa play for before the current team?	José Ernesto Sosa plays for the team Barcelona	Los Angeles Dodgers	América	José Ernesto Sosa played for the Chicago Cubs before the current team	Fenerbahçe SK
	Which employer did Nada Al-Nashif work for before the current employer?	Nada Al-Nashif worked for the current employer before 2021	Saudi Arabian Oil Company	World Bank	Nada Al-Nashif worked for the United Nations High Commissioner for Refuge	UNESCO
	Who is the head coach of the football club that José Sosa play for currently?	José Ernesto Sosa currently plays for the San Francisco Giants, coached by Gabe Kapler	José Luis ngel de la Rosa	José Ernesto Sosa currently plays for Club Atlético River Plate, and his head coach is Marcelo Gallardo	José Ernesto Sosa is a former Argentine footballer who played as a	Eduardo Domínguez
	Who is the head coach of the team that Oleksandr Filippov played for before the current team?	Oleksandr Filippov currently plays for Dynamo Kyiv	Oleksandr Yakovlev	Oleksandr Filippov does not play for any team currently, as he is a retired professional footballer	Oleksandr Filippov is a Ukrainian professional footballer who plays as	Bernd Hollerbach
	Who is the chief executive officer of the current employer of Ken Murphy?	Ken Murphy is the CEO of the current employer of himself	John Polaroid	Ken Murphy is the CEO of Visa	The current employer of Ken Murph is the CEO of the company	Ken Murphy
	Who is the spouse of the current head of the state of Indonesia?	The spouse of the current head of the state of Indonesia is Anak Jawaherrahimin	Susilo Bambang Yudhoyono	Iriana Widjaja	The current head of the state of Indonesia is Joko Widodo.	Iriana
	Who is the chair of the previous employer of Nada Al-Nashif?	Nada Al-Nashif is the current chair of the previous employer, Al-Nashif Group	Sheikha Moza bint Nasser Al-Nahyan	Nada Al-Nashif's previous employer is not specified in the article	Nada Al-Nashif is the chair of the previous employer of Nada Al	Audrey Azoulay

Figure 10: Responses of LLMs when prompted with questions prepending "Assume it is not December 2021," to the prompt.

Question	GPT-4 Web Answer	New Bing Answer	WikiData Gold Answer
Who is the head coach of the football club that José Sosa play for currently?	I couldn't find the current head coach for Estudiantes de La Plata directly from the sources reviewed. For the most accurate and up-to-date information regarding the head coach of Estudiantes de La Plata, where José Sosa currently plays, I recommend checking the official Estudiantes de La Plata website or sports news websites that cover Argentine football	José Sosa, the Argentine professional footballer, currently plays for FC Juárez. The head coach of FC Juárez is Afonso Sosa. If you have any more questions, feel free to ask!	Eduardo Domínguez
Who is the head coach of the team that Oleksandr Filippov played for before the current team?	Before joining his current team, Oleksandr Filippov played for Sint-Truiden in the Belgian First Division A during the 2020–21 season. To find out the head coach of Sint-Truiden during that time, you would need to look up the historical coaching records for the club around the 2020–21 season.	Oleksandr Filippov, the Ukrainian professional footballer, previously played for Desna Chernihiv before joining his current team, Dnipro-1. The head coach of Desna Chernihiv during Filippov's time there is not specified in the provided information. If you have any more questions, feel free to ask!	Bernd Hollerbach
Who is the chief executive officer of the current employer of Ken Murphy?	Ken Murphy is actually the Chief Executive Officer (CEO) of Tesco since October 2020, not someone employed by another organization looking for their CEO's information.	Ken Murphy himself is the CEO of his current employer, Tesco.	Ken Murphy
Who is the spouse of the current head of state of Indonesia?	The spouse of the current head of state of Indonesia, President Joko Widodo, is Iriana	The current President of Indonesia, Joko Widodo, is married to Iriana Jokowi. She has been the First Lady of Indonesia since October 20, 2014.	Iriana
Who is the chair of the previous employer of Nada Al-Nashif ?	Before her current role, Nada Al-Nashif worked at UNESCO as the Assistant Director-General for Social and Human Sciences from 2015 to 2019. To find the chairperson of UNESCO during her tenure, you would need to look up the leadership or executive board of UNESCO for that period	The previous employer of Nada Al-Nashif was UNESCO, where she served as the Assistant Director-General for Social and Human Sciences from 2015 until 201912. Unfortunately, the specific chairperson during her tenure at UNESCO is not mentioned in the provided information.	Audrey Azoulay

Figure 11: BingChat and GPT responses along with the gold answer for 5 sample multi-hop questions of our dataset. Correct answers are marked in green, outdated answers are marked in blue, correct one-hop answers but unable to find the two-hop answers are marked in orange and completely false or irrelevant answers are marked in red.