# ComplexTempQA: A 100m Dataset for Complex Temporal Question Answering

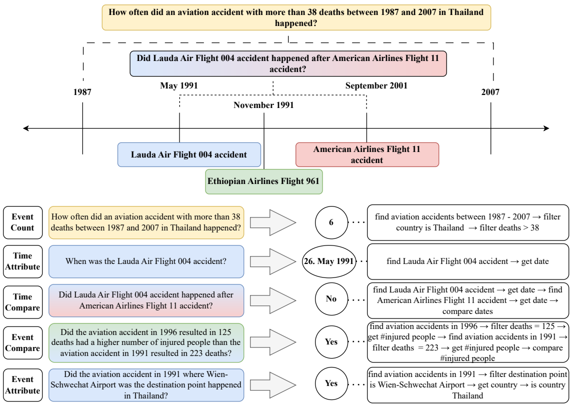
**Raphael Gruber[1], Abdelrahman Abdallah[1], Michael Färber[2], Adam Jatowt[1]**
[1]University of Innsbruck, [2]Technical University Dresden
{ra.gruber, abdelrahman.abdallah, adam.jatowt}@uibk.ac.at,
michael.faerber@tu-dresden.de

## Abstract

We introduce ComplexTempQA,[1] a large-scale dataset consisting of over 100 million question-answer pairs designed to tackle the challenges in temporal question answering. ComplexTempQA significantly surpasses existing benchmarks in scale and scope. Utilizing Wikipedia and Wikidata, the dataset covers questions spanning over two decades and offers an unmatched scale. We introduce a new taxonomy that categorizes questions as *attributes*, *comparisons*, and *counting* questions, revolving around events, entities, and time periods, respectively. A standout feature of ComplexTempQA is the high complexity of its questions, which demand reasoning capabilities for answering such as across-time comparison, temporal aggregation, and multi-hop reasoning involving temporal event ordering and entity recognition. Additionally, each question is accompanied by detailed metadata, including specific time scopes, allowing for comprehensive evaluation of temporal reasoning abilities of large language models.

## 1 Introduction

Temporal Question Answering (TQA) refers to answering questions that require both understanding of and reasoning about temporal knowledge (Jatowt, 2022; Jia et al., 2018b, 2019; Ning et al., 2020). This sets TQA apart from traditional Question Answering (QA). Developing effective TQA solutions naturally requires effective and challenging datasets. Existing TQA datasets, such as TORQUE (Ning et al., 2020), TEQUILA (Jia et al., 2019), ArchivalQA (Wang et al., 2021), and ChroniclingAmericaQA (Piryani et al., 2024), fall however short in several respects: *First*, they are limited in size, typically containing only a few thousand questions. This poses significant challenges for effectively training LLMs, as illustrated in (Ning et al., 2020; Jia et al., 2018a; Llorens et al., 2015; Ong et al., 2023; Wei et al., 2023; Naik et al., 2019). *Second*, while these datasets predominantly focus on specific question categories related to entities or time periods, they lack comprehensive coverage of a wide variety of question types. *Third*, they generally include only straightforward questions and omit complex inquiries that require multi-step inference to generate accurate responses, thereby limiting the ability of trained models to handle temporal reasoning tasks (Ning et al., 2020; Ong et al., 2023; Wei et al., 2023; Naik et al., 2019). *Fourth*, the prior datasets lack features such as popularity scores, which could indicate the relative ease of answering questions. They also do not allow filtering by specific time periods, a capability essential for detailed and customized temporal studies.


How often did an aviation accident with more than 38 deaths between 1987 and 2007 in Thailand happened?

Did Lauda Air Flight 004 accident happened after American Airlines Flight 11 accident?

How often did an aviation accident with more than 38 deaths between 1987 and 2007 in Thailand happened?

find aviation accidents between 1987 - 2007 → filter country is Thailand → filter deaths > 38

When was the Lauda Air Flight 004 accident?

find Lauda Air Flight 004 accident → get date

Did Lauda Air Flight 004 accident happened after American Airlines Flight 11 accident?

find Lauda Air Flight 004 accident → get date → find American Airlines Flight 11 accident → get date → compare dates

Did the aviation accident in 1996 resulted in 125 deaths had a higher number of injured people than the aviation accident in 1991 resulted in 223 deaths?

find aviation accidents in 1996 → filter deaths = 125 → get #injured people → find aviation accidents in 1991 → filter deaths = 223 → get #injured people → compare #injured people

Did the aviation accident in 1991 where Wien-Schwechat Airport was the destination point happened in Thailand?

find aviation accidents in 1991 → filter destination point is Wien-Schwechat Airport → get country → is country Thailand

Figure 1: Example types of temporal reasoning in questions sampled from ComplexTempQA (left) with the required inference steps (right) and timeline-based visualization (above).

To address these challenges, we introduce ComplexTempQA, a novel dataset that surpasses existing resources in both scale and complexity. It is particularly suitable for the analysis, training, and evaluation of LLMs and QA systems on com-

[1] Dataset and code available at: https://github.com/DataScienceUIBK/ComplexTempQA

*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9100–9112
November 4-9, 2025 ©2025 Association for Computational Linguistics

plex temporal knowledge. Our dataset offers four key contributions:

1. *Scale:* ComplexTempQA comprises over *100 million question-answer pairs*, making it by far the largest dataset available for Temporal QA.

2. *Question Types and Taxonomy:* The dataset includes diverse categories of questions, such as *attribute*-, *comparison*-, and *counting*-type questions, each pertaining to events, entities, or time periods. Questions are generated at scale based on facts from Wikidata and representative question patterns identified from Wikipedia, ensuring broad coverage of domains and question types. Moreover, we employ the IPTC Media Topics taxonomy to diversify thematic scope.

3. *Complexity and Temporal Scope:* Questions in ComplexTempQA require advanced reasoning skills, including event/entity matching, multi-hop inference, cross-time comparisons, and temporal ordering.

4. *Metadata and Evaluation:* Each question is enriched with detailed metadata, including the relevant time period within the overall dataset's time span from 1987 to 2023 and a popularity score that categorizes questions as popular or unpopular based on the type of question and the anticipated familiarity with the entities involved. This metadata allows for precise training and evaluation of language models, specifically concerning their ability to adapt to varying temporal contexts over time.

ComplexTempQA serves multiple purposes in advancing the study and application of LLMs in reasoning over temporal factual knowledge. Primarily, it enables a thorough analysis of LLM performance by addressing the need to understand temporal factual knowledge, identify temporal blind spots, and assess temporal reasoning capabilities (Wallat et al., 2024; Wenzel and Jatowt, 2023). Beyond performance evaluation, the dataset provides a foundational platform for the development of advanced question-generation tools. Our structured taxonomy facilitates the refinement and creation of specialized taxonomies for targeted tasks and datasets. Moreover, specific subsets of ComplexTempQA can be filtered to focus on particular types of temporal questions or particular desired time frames, supporting targeted research.

A key component of our work is evaluating how state-of-the-art LLMs handle the complex temporal questions posed by ComplexTempQA. In Section 6, we benchmark a range of models using various approaches, including zero-shot, few-shot, and retrieval-augmented generation (RAG). These evaluations offer insights into the current capabilities and limitations of LLMs in processing temporal information.

Overall, we make the following contributions:

1. We introduce and publicly release ComplexTempQA[1], a large-scale dataset comprising over 100 million question-answer pairs for temporal question answering, structured around a novel taxonomy of temporal question types.

2. We detail our methodology for dataset creation, which includes data retrieval from Wikidata, event extraction from Wikipedia, rigorous complexity assessment and rating, as well as filtering based on temporal ambiguity. Moreover, the dataset can be easily extended to cover additional time periods.

3. We benchmark diverse LLMs on ComplexTempQA using zero-shot, few-shot prompting, and retrieval-augmented generation (RAG) approaches, providing insights into their performance on temporal question answering tasks of varying complexity.

## 2 Related Work

Table 1 gives an overview of question answering datasets showing a notable discrepancy in question volume, with our dataset substantially surpassing others. While several TQA datasets contain complex temporal questions, they largely revolve around *TimeAttr* (time attribute) inquiries, which focus on relatively simple questions such as ones about the duration and order of events or entities. TORQUE (Ning et al., 2020) introduces questions that emphasize temporal relations such as "before," "after," and "start." The authors trained a model to evaluate questions specifically based on these temporal constraints. QA TempEval dataset (Llorens et al., 2015) has been designed with a focus on temporal entities and relations, which are easier to generate automatically. TEQUILA (Jia et al., 2019) uses temporal expressions like dates or implicit temporal signals such as "before" or "after."

In addition to these works, TempQuestions (Jia et al., 2018b) has been released as a benchmark for

| Dataset | Temporal Questions | #Questions | Creation Method | Source | Answer Type | Complex Question Types | Question Type | Time Frame | Temporal Metadata | Multi-Hop |
|---|---|---|---|---|---|---|---|---|---|---|
| NewsQuizQA (Lelkes et al., 2021) | No | 20K | Crowd sourced | News | Multiple choice | - | Attr | 2018–2022 | No | No |
| NewsQA (Trischler et al., 2016) | Partially | 119K | Crowd sourced | News | Extractive | TimeAttr | Attr | 2007–2015 | No | No |
| HOTPOTQA (Yang et al., 2018) | No | 113K | Crowd sourced | Wikipedia | Extractive | - | Attr, Comp | Unspecified | No | 2 hops |
| LC-QuAD 2.0 (Dubey et al., 2019) | Partially | 30K | Crowd sourced | Wikipedia | Extractive | TimeAttr | Attr, Count | Unspecified | No | No |
| TORQUE (Ning et al., 2020) | Yes | 21K | Crowd sourced | News | Generative | TimeAttr | Attr | Unspecified (short) | No | No |
| Time-Sensitive-QA (Chen et al., 2021) | Yes | 41K | Aut. Generated | Wikipedia | Extractive | TimeAttr | Attr | Unspecified (long) | No | No |
| TempQuestions (Jia et al., 2018a) | Yes | 1K | Aut. Generated | Freebase | Extractive | TimeAttr,<br>TimeComp,<br>TimeCount | Attr, Comp, Count | History | No | No |
| TKGQA (Ong et al., 2023) | Yes | 5K | Aut. Generated | News | Extractive | TimeAttr | Attr | 2022 | No | No |
| MenatQA (Wei et al., 2023) | Yes | 2K | Aut. Generated | Wikipedia | Extractive | TimeAttr | Attr | Unspecified (long) | No | No |
| TDDiscourse (Naik et al., 2019) | Yes | 6K | Aut. Generated | News | Extractive | TimeAttr | Attr | Unspecified (short) | No | No |
| ArchivalQA (Wang et al., 2021) | Partially | 532K | Aut. Generated | News | Extractive | TimeAttr | Attr, Count | 1987–2007 | No | No |
| **COMPLEXTEMPQA** | **Yes** | **100,228K** | **Aut. Generated** | **Wikipedia** | **Extractive, Boolean** | **TimeAttr,<br>TimeComp,<br>TimeCount,<br>Unnamed questions** | **Attr, Comp, Count** | **1987–2023** | **Yes** | **≤ 2 hops** |

Table 1: Comparison of COMPLEXTEMPQA with existing datasets highlighting the key aspects of question creation methodologies, answer types, complexity, temporal scope, and structure. Attr means *attribute*-type question, Comp denotes *comparison*-type questions and Count are *counting*-type questions.

temporal questions, containing 1,271 questions that are all temporal in nature, paired with their answers. That work provides a simple definition for temporal questions and demonstrates the need for further research on complex queries. Stricker (Stricker, 2023) applies answer extraction techniques from general question answering to retrieve temporal answers by identifying and processing time expressions. Their approach focuses on structured temporal information and distinguishes between absolute and relative time expressions.

Unlike other datasets, ours stands out due to its unique characteristics: **(a)** it comprises a number of question-answer pairs that is orders of magnitude larger than those in other datasets, **(b)** it categorizes these questions into specific types, **(c)** it includes complex questions, and **(d)** the questions are temporal in nature, with each strictly assigned to a specific time span.

## 3 Dataset Characteristics

We describe the COMPLEXTEMPQA characteristics along the four dimensions.

**Size:** COMPLEXTEMPQA comprises 100 million question-answer pairs and covers the period from 1987 to 2023. The dataset has been curated to probe the understanding of temporal knowledge within this 36-year span, encapsulating events, entity milestones, and other time-sensitive data.

**Question Types and Taxonomy:** The primary objective when constructing COMPLEXTEMPQA was to incorporate a broad spectrum of temporal questions. Our dataset offers extensive coverage across diverse subjects by aligning with IPTC MediaTopic standards [2], ensuring a broad range of relevant topics. The questions are also organized using a taxonomy that categorizes them by the questioned entity or event and the nature of the knowledge they probe, divided into *attribute*, *comparison*, and *counting* questions (see Figure 2). The numbers for each question type are shown in Table 2. Below we provide description of each type, while specific examples are given in Appendix A.

| Name | Total |
|---|---|
| Attribute Event | 83,798 |
| Attribute Entity | 84,079 |
| Attribute Time | 9,454 |
| Comparison Event | 25,353,340 |
| Comparison Entity | 74,678,117 |
| Comparison Time | 54,022,952 |
| Counting Event | 18,325 |
| Counting Entity | 10,798 |
| Counting Time | 12,732 |
| Multi-Hop: | 76,933 |
| Unnamed Event: | 8,707,123 |
| **Total:** | **100,228,457** |

Table 2: Dataset distribution (time questions are integrated within events or entities).

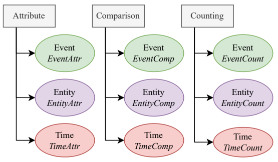


Figure 2: Taxonomy of temporal question types.

*Attribute*-type questions ask about the properties of events or entities, or relate to a specific time (e.g., “*When was the fall of the Berlin Wall?*”). Answers usually consist of **events**, **entities**, **dates** (or lists of dates), or **real numbers**. For numerical answers involving units, we convert all measurements to the International System of Units (SI units) to ensure

[2] https://iptc.org/standards/media-topics/

consistency.

*Comparison*-type questions involve the comparison of up to three events, entities, or time periods. The comparison aspect can be either numerical or temporal. For example, we compare two temporal attributes in the question: *"Did Halabja chemical attack happen after John F. Kennedy Jr. plane crash?"*. The answer to these questions can be either **true**, **false**, an **entity**, or an **event**. Due to the large number of possible questions that can be created based on comparison, this type is the most frequent in our dataset. We note that comparative questions require more reasoning than relatively simpler attribute type questions.

*Counting*-type questions—asking about a form of aggregation—record the frequency of a particular event type or the occurrence of an attribute for an entity. An example is *"How often did an aviation accident with more than 38 deaths occur in Thailand between 1987 and 2023?"* To make the answering of *counting*-type questions more precise, we always consider a specific time period and, in some cases, an attribute threshold. The threshold is determined by calculating the average value of the attribute across all comparable instances. For example, in the given example, the threshold refers to the number of deaths associated with the event. The answer to these questions is always a **natural number**.

**Popularity of Questions.** Questions are categorized as either **popular** or **unpopular** based on a popularity score derived from the English Wikipedia's page view statistics and the intrinsic complexity of the questions (see Sec. 4.1). A question is considered *popular* if all of its constituent entities or events are rated as common—that is, they have high page view counts that reflect broad public familiarity. We use the standard deviation for thresholds since page views follow a long-tail distribution. An average-based threshold would misclassify low-view questions as well-known. This approach helps ensure that notable events and entities are classified as popular. For example, a popular *time attribute*-type question is: *"When was the death of Diana, Princess of Wales?"*

Questions that require advanced reasoning—such as *counting*-type questions, multi-hop queries, or modified versions of standard questions that implicitly reference an event without explicitly naming it (which we refer to as **unnamed event questions**; see Sec. 4.2 for details)—are inherently more challenging and are automatically classified as *unpopular*.

**Metadata and Evaluation.** COMPLEXTEMPQA includes several metadata fields for each question, serving multiple purposes such as facilitating the retrieval of entities and events from the question or answer, enabling in-depth analysis of the dataset, and supporting segmentation based on attributes such as type, year, or popularity score. Specifically, the metadata comprises the following:

- **Type of Question:** The question type is specified based on the taxonomy shown in Figure 2.
- **Identifiers:** All corresponding Wikidata item identifiers for the entities or events that are either the subject of the question or part of the answer are included.
- **Country:** The country associated with the questioned entities or events and their answers is provided, if applicable.
- **Complexity Indicators:** These indicate properties used for generating multi-hop questions and whether the question is an unnamed event question.
- **Popularity Rating:** The popularity of the question is rated based on whether it concerns popular or unpopular entities or events.
- **Time Span:** A temporal range is specified for each question.

Appendix B provides further details on each metadata field together with an illustrative example.

## 4 Dataset Creation Pipeline

In this section, we present our methodology for constructing the dataset. We first describe the data acquisition process from Wikipedia and Wikidata, followed by the procedure for generating multi-hop questions. Finally, we detail the formation of question-answer pairs. An overview of the entire pipeline is illustrated in Figure 3.

### 4.1 Dataset Source Extraction

Different question types call for distinct data sources to best capture the required information while ensuring reliability. For the *event attribute*-type questions, we extracted every entry from "single year" Wikipedia pages (see, for example, year 1987[3]), collecting information on significant events

[3] https://en.wikipedia.org/wiki/1987

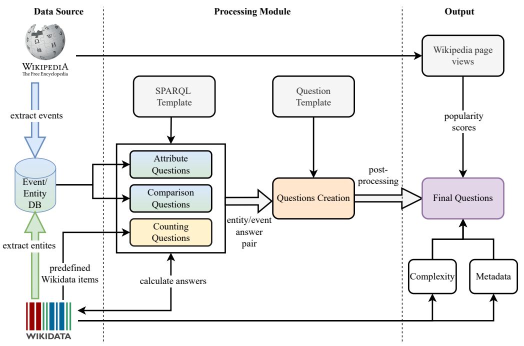


Figure 3: A simplified view of the dataset creation pipeline, showing how events and entities from Wikipedia and Wikidata are processed into final question–answer pairs with metadata and popularity scores.

occurring over a 36-year span (1987–2023). We then applied a filtering step to discard entries lacking a clear timestamp or falling outside our time frame.

We used the same Wikipedia sources to generate *comparison*-type questions, comparing up to three events (or entities) by examining attributes such as date, location, or other numerical values. These questions typically include terms like "higher/lower," "before/after," or "happened first."

To construct *entity attribute*-type questions, we compiled a list of entities by querying Wikidata through SPARQL, focusing on items such as "movies" and "heads of state." These predefined items were derived from the IPTC MediaTopic standards and manually curated. Because each item can yield an extensive list of possible entities, we filtered results based on the Wikipedia page views of those entities. For example, a raw list of over 100,000 "movie" entities was reduced to 35,991 by applying a page-view threshold.

Lastly, for *counting*-type questions, we used a distinct approach, as these questions enumerate sets of events or entities instead of focusing on a single event or entity. We began by creating SPARQL query templates tailored to specific counting tasks. For instance, by querying the Wikidata item "Presidents of the United States" within a specific time range, we obtained a list of all relevant individuals for potential *counting*-type questions. In some cases, we could directly use the resulting list; however, certain lists might be incomplete if Wikidata omits minor events (e.g. earthquakes). To mitigate such omissions, we introduced a threshold. For example, for "number of deaths in some calamity," we used the average values of corresponding attributes from Wikidata to define the cutoff.

## 4.2 Dataset Complexity Enhancement

We next increased the complexity of the dataset by introducing a module to create *multi-hop questions*. The idea is to leverage shared properties across the events and entities extracted in the previous step, requiring a multi-step reasoning processes to answer a question.

**Example of Multi-Hop Question.** Consider the question:

> *"What was the highest point of the country where the 1988 Summer Olympics happened, in meters?"*

1. Select a specific event (the *1988 Summer Olympics*) as the initial anchor.
2. Retrieve a property from that event (*country*).
3. Formulate a further query about that country by selecting an additional attribute (the *highest point*, in this case).

This multi-hop process requires multiple pieces of information across different domains—first about an event (location, date), then about a geographical feature related to that location. Multiple hops necessitate more complex computation which can be especially challenging in temporal settings when asking about more obscure events or entities from the past.

As another complexity enhancement, we created additional event references with implicit naming to expand the range of questions. For instance, "Lauda Air Flight 004 accident" was rephrased as "the aviation accident in 1991 which resulted in 223 deaths." We formed such references by combining the event's year with a property (e.g., number of fatalities), then confirming via a SPARQL query that no other events shared these attributes to ensure the lack of temporal ambiguity.

## 4.3 Dataset Formation and Enrichment

The final step involved constructing the actual question–answer pairs by integrating events, entities, and counting results into predefined question templates (see Appendix A). Designing the templates for generating question-answer pairs presented several challenges. The templates needed to be general enough to cover a broad range of temporal reasoning tasks while remaining structured enough to maintain coherence and logical validity.

The dataset had to support diverse temporal expressions, such as absolute dates, relative references, and temporal signals, requiring precise formatting and adaptability. Ensuring the generated questions were grammatically correct and naturally phrased was crucial, necessitating careful design to avoid awkward sentence constructions. Some questions required multiple reasoning steps involving different events and entities, making it challenging to construct templates that preserved logical consistency while maintaining clarity.

Below is an overview:

- *Attribute-type questions*: Query specific attributes of an event or entity. Following (Chen et al., 2021), we also included additional relations to enhance the *comprehensiveness* of these queries.

- *Comparison-type questions*: Select up to three events or entities to compare, using terms such as "higher/lower,", "smallest/largest," "before/after," or "happened first."

- *Counting-type questions*: Specify a time frame (e.g. five years) and query the relevant category (e.g. earthquakes, Nobel Peace Prize recipients) to count the number of matching events or entities.

Once the QA pairs were generated, we enriched them with *metadata* such as corresponding Wikidata IDs, the relevant country, and the specific time frame of the question. We additionally assigned *popularity scores* as described in Section 3.

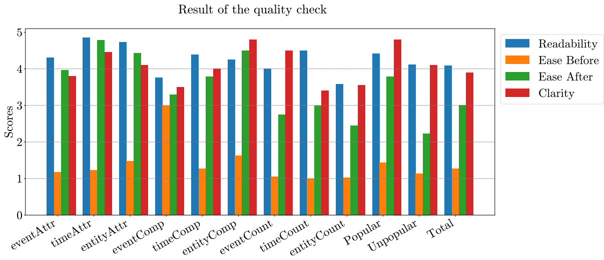


Figure 4: Result of the quality check.

| Method | Parameters | Precision | Recall | F1 | Con |
|---|---|---|---|---|---|
| Zephyr | 7B | 3.76 | 33.50 | 4.90 | 55.97 |
| Falcon | 7B | 0.31 | 34.22 | 0.62 | **64.21** |
| Llama-chat 7B | 7B | 3.68 | 33.94 | 6.09 | 50.32 |
| Mistral | 7B | 3.73 | **48.34** | 6.33 | 58.31 |
| LLama-chat 13B | 13B | 3.61 | 32.77 | 6.05 | 53.27 |
| Vicuna | 33B | 2.02 | 37.27 | 3.63 | 58.26 |
| Mixtral | 8x7B | 3.34 | 40.41 | 5.65 | 60.21 |
| LLama-chat 70B | 70B | **5.19** | 39.30 | 8.31 | 57.33 |
| Wizardlm | 70b | 1.63 | 27.63 | 2.80 | 50.44 |
| GPT-3.5 | - | 2.68 | 29.12 | 4.56 | 46.79 |
| GPT-4-o | - | **8.30** | 29.90 | **10.47** | 46.55 |

Table 3: Performance of zero-shot LLMs.

## 5 Dataset Quality Assessment

The quality assessment by human raters is essential to ensure that COMPLEXTEMPQA meets the high standards of clarity and readability required for advanced research. Expert evaluation allows us to detect and address potential errors, ambiguities, or biases, thereby reinforcing the dataset's reliability.

To evaluate COMPLEXTEMPQA, we conducted a user study with 11 volunteers (4 females, 7 males; 4 with secondary education and 7 with postgraduate degrees; ages 26–56, predominantly in their 20s and 30s). Participants assessed 450 randomly selected questions, evenly distributed across all types, using a 5-point Likert scale across four dimensions: (1) **readability**; (2) **ease of answering** *prior to using a web search engine*; (3) **ease of answering** *after conducting a web search*; and (4) overall **clarity** (including lack of ambiguity). For dimensions (2) and (3), participants first rated the questions based solely on their own knowledge (2), and then again after performing web search (3) with explicit instructions not to use any large language models.

Figure 4 shows that questions generally received high ratings for readability and clarity. However, participants reported that answering questions without a search engine was challenging, while the access to web search results significantly improved answerability. Notably, *counting*-type questions remained the most difficult to answer even after web searches, as they often require special type of inference. Additionally, some *counting*- and *comparison*-type questions received lower clarity scores, suggesting opportunities for further refinement. Overall, the expected familiarity of the questions significantly influenced the ratings, with questions classified as unpopular proving more challenging both before and after web search.

## 6 Experiments

We evaluate multiple Large Language Models (LLMs), including Llama, Mistral, Mixtral, Falcon, Vicuna, Zephyr, WizardLM, GPT-3.5, and GPT-4-o, each selected for its unique strengths. We discuss the models in Appendix C, while Appendix D lists the prompts used in experiments.

Evaluating LLMs (Guo et al., 2023; Abdallah et al., 2023), especially for question answering, is challenging due to the verbose nature of the re-

| Method | Parameters | Shots | Precision | Recall | F1 | Con | EM |
|---|---|---|---|---|---|---|---|
| Llama-2 | 7B | 0 | 3.675 | **33.935** | 6.09 | **50.315** | 0.035 |
| | | 1 | 7.08 | 27.21 | 9.00 | 43.09 | 3.57 |
| | | 2 | 23.05 | 30.65 | 23.655 | 33.06 | 21.49 |
| | | 3 | **23.67** | 30.38 | **24.22** | 31.78 | **22.25** |
| Llama-2 | 13B | 0 | 3.605 | **32.76** | 6.05 | **53.27** | 0.0085 |
| | | 1 | 22.865 | 27.91 | 23.345 | 38.11 | 21.525 |
| | | 2 | **31.645** | 32.35 | **31.56** | 31.465 | **30.37** |
| | | 3 | 30.37 | 32.03 | 30.495 | 32.46 | 29.00 |
| Llama-2 | 70B | 0 | 5.191 | 39.30 | 8.31 | **57.32** | 0.138 |
| | | 1 | 25.74 | 32.61 | 18.8 | 49.60 | 13.61 |
| | | 2 | 34.01 | 44.21 | 35.26 | 43.27 | 31.78 |
| | | 3 | **37.09** | **46.57** | **38.44** | 42.77 | **34.59** |
| Mistral-Instruct | 7B | 0 | 3.73 | **48.33** | 6.325 | **58.30** | 0.034 |
| | | 1 | 24.68 | 35.33 | 25.87 | 41.81 | 21.83 |
| | | 2 | 32.74 | 35.91 | 33.145 | 34.82 | 30.55 |
| | | 3 | **35.28** | 37.91 | **35.68** | 36.26 | 32.93 |
| Mixtral | 8x7B | 0 | 3.34 | 40.41 | 5.65 | **60.20** | 0.156 |
| | | 1 | 6.76 | 39.06 | 9.17 | 52.87 | 2.75 |
| | | 2 | 14.27 | 41.89 | 16.44 | 52.64 | 10.22 |
| | | 3 | **15.49** | **44.03** | **17.83** | 51.57 | **11.25** |
| GPT-3.5 | - | 0 | 2.68 | 29.12 | 4.56 | **46.79** | 0.008 |
| | | 1 | 21.12 | 45.58 | 23.95 | 52.77 | 15.80 |
| | | 2 | 30.88 | 41.17 | 32.71 | 37.79 | 26.01 |
| | | 3 | **31.68** | **42.37** | **33.40** | 38.85 | **26.53** |
| GPT-4o | - | 0 | 8.30 | 29.90 | 10.47 | 46.55 | 3.45 |
| | | 1 | 21.12 | 45.58 | 23.95 | 52.77 | 15.80 |
| | | 2 | 40.82 | 53.83 | 42.92 | 48.12 | 35.62 |
| | | 3 | **43.91** | **56.08** | **45.72** | **47.62** | **39.07** |

Table 4: Performance of few-shot LLM Models.

| Method | Parameters | Context | Precision | Recall | F1 | Con |
|---|---|---|---|---|---|---|
| Llama-2 | 7B | No Context | 3.67 | 33.93 | 6.09 | 50.31 |
| | | Retriever | 3.59 | 33.67 | 5.97 | 53.48 |
| | | True Context | **3.92** | **37.40** | **6.49** | **56.03** |
| Llama-2 | 13B | No Context | 3.60 | 32.76 | 6.05 | 53.27 |
| | | Retriever | 3.50 | 34.22 | 5.84 | 55.38 |
| | | True Context | **3.75** | **37.09** | **6.28** | **57.42** |
| Llama-2 | 70B | No Context | 5.19 | **39.30** | 8.31 | 57.32 |
| | | Retriever | 5.27 | 36.16 | 8.12 | 56.45 |
| | | True Context | **5.82** | 38.59 | **8.82** | **58.26** |
| Mistral-Instruct | 7B | No Context | 3.73 | **48.33** | 6.32 | **58.30** |
| | | Retriever | 3.86 | 33.32 | 6.31 | 55.90 |
| | | True Context | **5.13** | 35.26 | **8.08** | 54.14 |
| Mixtral | 8x7B | No Context | 3.34 | **40.41** | 5.65 | 60.20 |
| | | Retriever | **4.23** | 35.93 | **6.62** | 56.30 |
| | | True Context | 3.65 | 38.02 | 5.88 | **60.54** |

Table 5: Performance of LLMs in RAG QA setting.

sponses. Traditional metrics like Exact Match and F1 score may not be suitable. To address this, we use model-agnostic metrics like Token Recall and Answer String Containment[4]. Token Recall measures how well the model's response covers the ground truth. Answer String Containment assesses if the model's response captures the core answer.

## 6.1 Zero-shot Results

We conducted zero-shot QA experiments to evaluate different Large Language Models (LLMs). The models generate responses based solely on their pre-training. The results, presented in Table 3, show varying performance across different metrics. Notably, model size is not the sole determinant of performance. For instance, Llama-chat models, with fewer parameters, perform comparably to GPT-3.5. Some models, like GPT-4o, Vicuna and Mistral, suggest a trade-off between precision and comprehensiveness. Models like Zephyr and Falcon, despite lower precision and F1 scores, have high recall and containment scores, indicating their ability to capture significant portions of the ground truth. Lastly, the WizardLM model was found to have lower scores across all metrics.

## 6.2 Few-shot Results

In the few-shot learning setting (Chada and Natarajan, 2021), models improve as they are provided with more examples, as seen in Table 4. Across all models, performance increases with additional shots, but the rate of improvement plateaus after two shots, indicating diminishing returns. The Llama-2 7B, 13B, and 70B models exhibit steady gains, with the 70B variant achieving the highest performance among them. Similarly, Mistral-Instruct and Mixtral models follow the same trend, though with smaller absolute gains. Notably, GPT-4o outperforms all models, showing a significant improvement in F1 score from 10.47 (zero-shot) to 45.72 (three-shot), along with the highest recall and containment scores, demonstrating its superior ability to adapt with few-shot examples. Finally, as expected, GPT-3.5's performance is much worse than the one of GPT-4o.

## 6.3 RAG Results

Retrieval-augmented generation (RAG) (Lewis et al., 2020; Abdallah and Jatowt, 2023) combines the strengths of pre-trained language models and information retrieval systems to generate responses in a question-answering setting. In RAG, when a question is posed, relevant documents are first retrieved from a large corpus. These retrieved documents are then provided as additional context to a language model, which generates a response based on both the original question and the retrieved documents. Following (Karpukhin et al., 2020), we use the English Wikipedia dump from Dec. 20, 2018 as the source documents for answering questions, which contains 21,015,324 passages in total. Each passage is prepended with the title of the Wikipedia article from which it originates, along with an [SEP] token.

In this experiment, we aimed to evaluate the efficiency of appending the retrieved context to LLMs retrieved by Dense Passage Retriever (DPR) (Karpukhin et al., 2020) for question answering. We tested three different settings for each

[4] https://huggingface.co/spaces/evaluate-metric/squad

| Method | Parameters | Attribute-type | | | | Comparison-type | | | | Counting-type | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Con | Precision | Recall | F1 | Con | Precision | Recall | F1 | Con |
| Llama-2 | 7B | 6.55 | 60.70 | 9.52 | 96.86 | 5.30 | 58.05 | 8.83 | 63.20 | 0.34 | 10.50 | 0.66 | 80.24 |
| Llama-2 | 13B | 6.17 | 61.26 | 10.67 | 96.77 | 5.22 | 59.53 | 8.75 | 63.30 | 0.16 | 5.86 | 0.30 | 86.13 |
| Llama-2 | 70B | 8.21 | 65.85 | 13.71 | **97.63** | 7.06 | 61.41 | 11.25 | 64.76 | 0.43 | 16.42 | 0.84 | 48.99 |
| Mistral-Instruct | 7B | 5.87 | 54.00 | 9.96 | 96.52 | 8.33 | 57.79 | 12.73 | 62.18 | 0.68 | 20.80 | 1.31 | 87.45 |
| Mixtral | 8*7B | 4.04 | **75.80** | 7.21 | 97.25 | 5.18 | 63.50 | 8.86 | **72.10** | 0.60 | **27.44** | 1.17 | **90.10** |
| GPT-3.5 | - | 9.48 | 73.16 | 15.87 | 70.02 | 6.51 | 55.07 | 10.41 | 43.50 | 0.20 | 4.84 | 0.39 | 77.38 |
| GPT-4o | - | **23.94** | 75.56 | **30.05** | 72.66 | **12.38** | **71.63** | **17.77** | 57.17 | **2.16** | 20.67 | **3.46** | 78.61 |

Table 6: Performance on Attribute-, Comparison-, and Counting-type questions.

| Method | Parameters | Entity | | | | Event | | | | Time | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Con | Precision | Recall | F1 | Con | Precision | Recall | F1 | Con |
| Llama-2 | 7B | 4.26 | 40.86 | 6.95 | 75.80 | 4.75 | 53.25 | 7.87 | 82.00 | 3.38 | 41.86 | 5.84 | 76.09 |
| Llama-2 | 13B | 3.79 | 40.70 | 6.52 | 80.95 | 4.96 | 49.72 | 8.31 | 81.80 | 3.14 | 38.27 | 5.45 | 81.28 |
| Llama-2 | 70B | 5.07 | 47.55 | 8.45 | 82.31 | 7.23 | 60.18 | 11.28 | 85.10 | 4.50 | 53.50 | 7.67 | 84.02 |
| Mistral-Instruct | 7B | 3.91 | 35.43 | 6.53 | 79.20 | 7.49 | **62.28** | 11.64 | 84.60 | 4.31 | 39.85 | 6.95 | 81.55 |
| Mixtral | 8*7B | 2.82 | 50.95 | 4.85 | **84.76** | 4.91 | 60.06 | 8.29 | **86.60** | 2.62 | 54.78 | 4.77 | **85.85** |
| GPT-3.5 | - | 5.59 | 43.64 | 9.32 | 64.47 | 6.80 | 52.17 | 10.73 | 58.63 | 4.33 | 39.23 | 7.32 | 64.63 |
| GPT-4o | - | **14.10** | **51.55** | **17.10** | 72.03 | **13.77** | 57.17 | **18.02** | 59.76 | **11.00** | **57.12** | **15.85** | 72.78 |

Table 7: Performance on Entity, Event, and Time questions.

model: *without context*, *with the first top retrieved passage* as context, and *with the true context*. The true context is determined by retrieving $1,000$ passages using DPR and conducting a simple search within these passages. If the answer was found within a passage, we selected the first passage that contains the answer as the true context. If the answer was not found in any of the retrieved passages, we selected a random passage.

Table 5 presents the results. The performance of the models usually improves when context is provided, with the true context generally leading to the best performance. This suggests that providing relevant context can help guide the models in generating more accurate and relevant responses. However, the performance varies across different models and settings, indicating that the effectiveness of RAG depends on both the specific model and the quality of the retrieved context.

### 6.4 Results on Different Question Types

Across different question types—attribute, comparison, counting, entity, event, and time—the performance of LLMs varies significantly, as shown in Tables 6 and 7. Counting type questions are most challenging followed by the comparison questions and then the attribute type questions.

Llama models exhibit a steady improvement with increasing parameters, with Llama-2 70B outperforming its smaller variants across most metrics. However, Mistral-Instruct and Mixtral, despite having fewer parameters, achieve comparable or better results in certain cases, particularly in recall and containment scores. GPT-4o consistently delivers the highest performance across all categories, achieving the best precision, recall, and F1 scores, particularly excelling in attribute and comparison-type questions. It also dominates in entity, event, and time-based questions, highlighting its strong generalization ability. GPT-3.5 performs well but falls behind GPT-4o, particularly in recall and containment, indicating a weaker ability to retrieve and structure temporal knowledge.

## 7 Conclusions

We introduced COMPLEXTEMPQA, a large-scale dataset comprising over 100 million temporal question-answer pairs, surpassing existing benchmarks in scope, coverage, and complexity. Built on Wikipedia and Wikidata, it spans more than 30 years and covers a wide range of domains, including history, politics, sports, and science. We introduced a taxonomy categorizing questions into three key types: *attributes*, *comparisons*, and *count*, each requiring advanced temporal knowledge and temporal reasoning skills such as multi-hop inference, temporal aggregation, and event ordering. To support targeted evaluation, each question is enriched with structured metadata, enabling precise assessment of LLMs' ability to process and reason over temporal information. In Appendix E we discuss the different use cases of our dataset.

Finally, we evaluated several LLMs, revealing significant gaps in their capabilities. While state-of-the-art models performed relatively well on simpler questions, their performance on more complex temporal questions decreased significantly, highlighting the challenging character of our dataset.

## Limitations

Despite its advantages, COMPLEXTEMPQA has several limitations. The dataset is built upon Wikipedia and Wikidata, which are characterized by relatively high precision but may suffer from lower recall, meaning that while available facts are generally accurate, relevant historical or domain-specific facts might be missing. The dataset is also constrained by its timeframe, as it primarily covers the period of 1987 until 2023, limiting its applicability to broader historical analysis. Additionally, the temporal scope of questions may not always align perfectly with evolving real-world knowledge, as both Wikipedia and Wikidata are continuously updated. This is also a challenge in our RAG analysis for which we employ the Wikipedia dump from 2018 which is however commonly used as a retrieval corpus in RAG studies; thus we have adapted the same setting in our experiments. Moreover, the dataset contains a significant proportion of comparative questions, which, while valuable for evaluating comparative reasoning over historical knowledge, may introduce a bias towards one form of temporal inference. Addressing these challenges could improve the dataset's utility in future iterations.

## Acknowledgments

The computational results presented here have been achieved (in part) using the LEO HPC infrastructure of the University of Innsbruck.

## References

Abdelrahman Abdallah and Adam Jatowt. 2023. Generator-retriever-generator: A novel approach to open-domain question answering. *arXiv preprint arXiv:2307.11278*.

Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal qa systems. *arXiv preprint arXiv:2304.06623*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. 2014. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):1–41.

Rakesh Chada and Pradeep Natarajan. 2021. Fewshotqa: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models. *arXiv preprint arXiv:2109.01951*.

Wenhu Chen, Xinyi Wang, and William Yang Wang. 2021. A dataset for answering time-sensitive questions. *CoRR*, abs/2108.06314.

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-qa: A dataset for event-centric question answering over knowledge graphs. *CoRR*, abs/2004.11861.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Mohnish Dubey, Debayan Banerjee, Abdelrahman Abdelkawi, and Jens Lehmann. 2019. Lc-quad 2.0: A large dataset for complex question answering over wikidata and dbpedia. In *Proceedings of the 18th International Semantic Web Conference*, ISWC'19, pages 69–78.

Michael Färber, David Lamprecht, Johan Krause, Linn Aung, and Peter Haase. 2023. Semopenalex: The scientific landscape in 26 billion RDF triples. *CoRR*, abs/2308.03671.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, and 1 others. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Kelvin Han and Claire Gardent. 2023. Generating and answering simple and complex questions from text and from knowledge graphs. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 285–304.

Nathan Hu, Eric Mitchell, Christopher D Manning, and Chelsea Finn. 2023. Meta-learning online adaptation of language models. *arXiv preprint arXiv:2305.15076*.

Adam Jatowt. 2022. Temporal question answering in news article collections. In *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022*, page 895. ACM.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018a. 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.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2019. TEQUILA: temporal question answering over knowledge bases. *CoRR*, abs/1908.03650.

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

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Ádám D. Lelkes, Vinh Q. Tran, and Cong Yu. 2021. Quiz-style question generation for news stories. *CoRR*, abs/2102.09094.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James F. Allen, and James Pustejovsky. 2015. Semeval-2015 task 5: QA tempeval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 792–800.

Aakanksha Naik, Luke Breitfeller, and Carolyn P. Rosé. 2019. Tddiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 239–249. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. *CoRR*, abs/2005.00242.

Ryan Ong, Jiahao Sun, Ovidiu Serban, and Yi-Ke Guo. 2023. TKGQA dataset: Using question answering to guide and validate the evolution of temporal knowledge graph. *Data*, 8(3):61.

Bhawna Piryani, Jamshid Mozafari, and Adam Jatowt. 2024. Chroniclingamericaqa: A large-scale question answering dataset based on historical american newspaper pages. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2038–2048. ACM.

Evan Sandhaus. 2008. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12):e26752.

Tarcísio Souza Costa, Simon Gottschalk, and Elena Demidova. 2020. Event-qa: A dataset for event-centric question answering over knowledge graphs. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 3157–3164.

Armand Stricker. 2023. Question answering in natural language: the special case of temporal expressions. *arXiv preprint arXiv:2311.14087*.

Jihoon Tack, Jaehyung Kim, Eric Mitchell, Jinwoo Shin, Yee Whye Teh, and Jonathan Richard Schwarz. 2024. Online adaptation of language models with a memory of amortized contexts. *arXiv preprint arXiv:2403.04317*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR*, abs/1611.09830.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, and 1 others. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Bastian Haarmann, Anastasia Krithara, Michael Röder, and Giulio Napolitano. 2017. 7th open challenge on question answering over linked data (qald-7). In *Semantic Web Challenges: 4th SemWebEval Challenge at ESWC 2017, Portoroz, Slovenia, May 28-June 1, 2017, Revised Selected Papers*, pages 59–69. Springer.

Jonas Wallat, Adam Jatowt, and Avishek Anand. 2024. Temporal blind spots in large language models. *CoRR*, abs/2401.12078.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2021. Archivalqa: A large-scale benchmark dataset for open domain question answering over archival news collections. *CoRR*, abs/2109.03438.

Jiexin Wang, Adam Jatowt, Masatoshi Yoshikawa, and Yi Cai. 2023. Bitimebert: Extending pre-trained language representations with bi-temporal information. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 812–821. ACM.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. Menatqa: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 1434–1447.

Georg Wenzel and Adam Jatowt. 2023. An overview of temporal commonsense reasoning and acquisition. *CoRR*, abs/2308.00002.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *CoRR*, abs/1809.09600.

## A Example Questions and Templates

In our question generation process, we employed a range of templates, each tailored to different contexts and requirements. Below, we present several selected question templates showcasing the diversity of data:

**Attribute Queries**

- `What was [ATTR] of [ENTITY]?`
  *Example: What was the genre of the movie* Border*?*

- `When was [PERSON] [ENTITY]?`
  *Example: When was Girija Prasad Koirala President of Nepal?*

- `When [V1] [V2] [ENTITY]?`
  *Example: When was the publication date of the movie in which Lou Diamond Phillips and Esai Morales acted?*

**Comparison Queries**

- `Comparing [P1] of [ENTITY1] and [ENTITY2], which one has a [COMPARE] [P2]?`
  *Example: Comparing the country of the company Google and the company Yandex, which one has a lower highest point?*

- `Did [EVENT1] have a [COMPARE] [ATTR] than [EVENT2]?`
  *Example: Did the car bombing in 1993 (6 deaths) have a higher death toll than the train wreck in 1989 (645 deaths)?*

- `Did [EVENT1] or [EVENT2] happen first?`
  *Example: Did Helios Airways Flight 522 accident or the 27th G8 summit happen first?*

- `Which one happened first, [EVENT1], [EVENT2], or [EVENT3]?`
  *Example: Which happened first, the 69th Academy Awards, the USS Cole bombing, or the Daegu subway fire?*

- `Did [EVENT1] happen [SIGNAL] [EVENT2]?`
  *Example: Did the Gulf War happen after Hurricane Hugo?*

**Counting Queries**

- `How many times did [PERSON] win [ENTITY] [TIME]?`
  *Example: How many times did Robert Richardson win the Academy Award for Best Cinematography between 1987 and 2007?*

- `How often did [EVENT] happen [TIME]?`
  *Example: How often did an aviation accident with more than 54 participants in Argentina happen in 1999?*

- `In how many years did [EVENT] [YEAR] happen?`
  *Example: In how many years did a aviation accident with more than 28 survivors between 2008 and 2023 by Xiamen Airlines happen?*

**Multi-hop Queries**

- `Did [V] [ENTITY1] [SIGNAL] [V] [ENTITY2]?`
  *Example: Did the publication of the movie with Lou Diamond Phillips and Esai Morales happen before the publication of* Resident Evil: Extinction*?*

- `What was the [P2] of the [P1] of [ENTITY] in meters?`
  *Example: What was the highest point (in meters) of the country where the Moscow theater hostage crisis happened?*

We employ placeholders to convey specific elements within the question templates. The placeholder [SIGNAL] denotes temporal relationships, such as ”before” or ”after,” indicating the sequence of events or entities. [ATTR] stands for attributes like the number of injured people, providing context or additional information related to the events or entities. [YEAR] referring either to a specific year or between two years e.g. *”in 1987”* or *”between 1987 and 2007”*. [COMPARE] signifies comparison relationships, such as ”higher” or ”lower,” enabling the comparison of attributes or characteristics between events, entities, or their attributes. Then we have [V] to denote various verbs like ”was” or ”happened”. [P] stands for properties used to create multi hop questions. Furthermore, we make a clear distinction between questions involving persons and those involving other entities, as the question structure may vary accordingly. Despite these differences, some question templates are versatile enough to accommodate both event and entity questions, ensuring flexibility and adaptability in our approach to question generation.

## B Metadata

We provide the following metadata:[5]

**Entity in question:** A list of Wikidata IDs of the question.

**Entity in answer:** A list of Wikidata IDs of the answer if it contains an entity.

**Country in question:** A list of country Wikidata IDs of the countries of the questioned entities.

**Country in answer:** A list of country Wikidata IDs of the countries of the entities in the answer.

**Hop property:** A list of Wikidata properties of the question if it contains a hop. If there are multiple hops, they are listed in the order of use.

**Rating:** A numerical rating where *0* is considered a popular (or easy) question, and *1* is considered a less popular or harder question according to the rating as described in Section 3.

**Is unnamed:** A numerical which is *1* if the question contains an implicitly described event and *0* otherwise.

**Type:** The type based on the taxonomy given in Figure 2.

**Time span:** The time frame to which the question relates to. For example, for the entity questions, the time frame ranges from born/creation to death/destruction. The start is always the earlier date, and the end is the latter.

Below is an example for a question including the metadata:

- **Question**: What was the highest point of the country where the 1988 Summer Olympics happened, in meters?
- **Answer**: [1950]
- **Entity in question**: [8470]
- **Entity in answer**: []
- **Country in question**: [884]
- **Country in answer**: []
- **Hop property**: [17, 610]
- **Rating**: 1
- **Is unnamed**: 0
- **Time span**: [1988-09-17, 1988-10-02]

## C Models used in Experiments

In experiments we test multiple Large Language Models, including Llama, Mistral, Mixtral, Falcon, Vicuna, Zephyr, WizardLM, GPT-3.5, and GPT-4-o, each selected for its unique strengths. Llama models, developed by Meta (Touvron et al., 2023), leverage reinforcement learning with human feedback (RLHF) for dialogue optimization. Mistral-7B and Mixtral (Sparse Mixture of Experts) outperform Llama-2 13B and 70B, respectively, in various benchmarks (Jiang et al., 2023). GPT-3.5 improves upon GPT-3 by reducing toxicity and enhancing contextual understanding (Brown et al., 2020). Falcon, optimized for inference, utilizes multi-query attention and FlashAttention (Dao et al., 2022). Vicuna fine-tunes Llama-2 using ShareGPT data, enhancing conversational capabilities, while Zephyr employs distilled supervised fine-tuning (dSFT) for improved task accuracy (Tunstall et al., 2023). Finally, WizardLM, based on Llama-2 13B, refines instruction-following abilities using an Evol-Instruct method (Xu et al., 2023).

## D Experiment Prompts

In all experiments, each LLM was prompted to function as a helpful assistant and deliver direct, concise answers. For models employing the Retrieval Augmented Generation (RAG) approach,

[5] For the Wikidata IDs we exclude the leading *’Q’* as well for the properties the leading *’P’*

the prompt included additional context from retrieved documents to enhance the response quality. This modification ensured that responses were informed by relevant background information.

For all the experiments involving LLMs, we used the following prompt:

> You are a helpful assistant. Provide direct and concise answer to the following question.
> Question: <question>.

In the case of RAG, the prompt is slightly modified to incorporate the additional context provided by the retrieved documents. The prompt used for RAG is:

> You are a helpful assistant. Using the context, provide direct and concise answers to the following question.
> Question: <question>.
> Context: <context>.

## E Dataset Use

We briefly list below the intended use cases of our dataset.

**LLM Evaluation and Training.** COMPLEXTEMPQA is an effective resource for evaluating LLMs, as demonstrated in our preliminary study (Sec. 6). As the largest QA dataset currently available, it provides an unparalleled foundation for analyzing LLM performance. It supports finetuning, prompt engineering, and the assessment of temporal question answering capabilities (Wallat et al., 2024). Notably, the dataset facilitates rigorous evaluation of truthfulness by offering unprecedented diversity and scale—critical factors for mitigating hallucinations. Moreover, its rich temporal metadata expands the scope of time-based QA research (Costa et al., 2020).

**Continual Learning and Adaptation of LLMs.** The detailed temporal annotations and extensive scale of COMPLEXTEMPQA make it useful for online adaptation and continual training approaches (Hu et al., 2023; Tack et al., 2024). With approximately 280k questions per year on average, it enables targeted experiments on temporal adaptation—vastly outperforming benchmarks like ArchivalQA (Wang et al., 2021), which offers only around 22k questions per year.

**RAG Systems.** COMPLEXTEMPQA can be used to train and evaluate Open Domain Question Answering models with historical news archives, such as the NYT Annotated Archive (Sandhaus, 2008) (1.8 million articles from 1987 to 2007). Given the NYT's international coverage and our focus on US events, most questions can be answered using this archive, making our dataset a complementary resource for temporal IR research (Wang et al., 2021; Campos et al., 2014; Wang et al., 2023).

**KGQA Systems.** Built from Wikidata and Wikipedia, COMPLEXTEMPQA is well-suited for Knowledge Graph Question Answering (Usbeck et al., 2017; Souza Costa et al., 2020). Its integration with large-scale knowledge graphs such as Wikidata and SemOpenAlex (Färber et al., 2023)—which contain billions of facts—enhances QA models' ability to explore complex temporal relationships and evaluate multi-hop reasoning (Han and Gardent, 2023).