HistoryBankQA: Multilingual Temporal Question Answering on Historical Events

Biswadip Mandal
biswadip.iitb@gmail.com

Anant Khandelwal
anantk@microsoft.com

Manish Gupta
gmanish@microsoft.com

Abstract

Temporal reasoning over historical events is vital for temporal NLP tasks such as event extraction, entity linking, question answering (QA), timeline summarization, event clustering, and natural language inference. However, benchmarks for evaluating large language models (LLMs) on temporal reasoning remain limited. Existing datasets are small, lack multilingual coverage, and focus on recent events. To address this, we introduce HistoryBank, a multilingual database of 10M+ historical events sourced from Wikipedia timelines and infoboxes. Our database provides unprecedented coverage in both historical depth and linguistic breadth with 10 languages. We also present a comprehensive benchmark covering 6 temporal QA tasks across all languages, evaluating models like LLaMA-3-8B, Mistral-7B, Gemma-2-9B, Qwen3-8B, and GPT4o. GPT-4o consistently performs best; Gemma-2 leads among smaller models. Our work offers a rich resource for advancing multilingual, temporally-aware language understanding of historical events. To support further research, we publicly release our code and datasets[1].

1 Introduction

Temporal reasoning is the ability to understand, represent, and manipulate time-related information. In historical contexts, it underpins many natural language processing (NLP) and knowledge-based applications, including event extraction, temporal question answering (QA), timeline generation, historical entity linking, event clustering, timeline summarization, temporal natural language inference (NLI), planning, and narrative comprehension. Accurate event extraction requires identifying and timestamping events, which demands a deep understanding of historical timelines. Temporal QA answers questions like "Who was the US president during WWII?", demanding reasoning over historical periods and context. Timeline generation organizes events chronologically to build coherent narratives, while historical entity linking disambiguates mentions like "King George" using temporal cues. Event clustering groups events by temporal proximity or shared context, aiding pattern discovery. Timeline summarization condenses complex narratives, such as social media histories (Song et al., 2024), into structured, time-ordered summaries. Temporal NLI identifies whether one event entails or contradicts another. Improving temporal NLI has broad practical implications for information synthesis and reasoning.

Benchmarking large language models (LLMs) for temporal reasoning faces two main challenges: (1) creating a large, multilingual historical event corpus that spans diverse temporal contexts across cultures and periods, and (2) designing a comprehensive benchmark that uses this corpus to evaluate tasks like temporal entailment, ordering, duration inference, and temporal QA through rich, diverse, and challenging scenarios.

To address the first challenge, prior work has extracted events from web sources (Wang et al., 2019), social media (Chang et al., 2020), and documents using heuristic prompts (Liu and Luo, 2024). Recent efforts target multilingual extraction from low-resource historical corpora like colonial newspaper ads (Borenstein et al., 2023), facing issues of linguistic variation, annotation scarcity, and domain shift. These advances yield dynamic, localized, sometimes multilingual event databases, but remain focused on ephemeral, contemporary, or socially driven content. We discuss detailed related work in Appendix A.

Historical events, spanning centuries and covering political, scientific, cultural, and social milestones, are underrepresented in current resources,

[1]Code available at https://github.com/mandalbiswadip/history-bank and data available at: https://drive.google.com/drive/folders/1vHudioDdI3EeYPbhYjKa0gimxaXvpxB2

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 474–496

Question Type		Question	Answer
FactQA		I'm looking for the date of the release of the Goodnight Punpun Omnibus 1. When was it released?	{'year': '2016', 'month': '03', 'day': '04'}
DurationQA		What was the time duration between the release of the single "Happy Happy" and the release of the single "Fake &True"?	start_date={'year': '2019', 'month': '06', 'day': '12'}, end_date={'year': '2019', 'month': '10', 'day': '18'}
RelationQA	Compare Duration	Which event span is of longer duration: "Agnes van Ardenne was born" to "Agnes van Ardenne started her earlier term as Member of the House of Representatives", or "Agnes van Ardenne ended her earlier term as Member of the House of Representatives" to "Agnes van Ardenne started her term as Member of the House of Representatives"?	"Agnes van Ardenne was born" to "Agnes van Ardenne started her earlier term as Member of the House of Representatives".
	Duration Diff	What is the difference in duration between "A new session of the Alabama Legislature began" to "The next election for the Alabama Senate is scheduled", and "Anthony Daniels was elected as House Minority Leader of Alabama" to "The last election for the Alabama Senate was held". (in days) in the context of "Alabama Legislature"?	762
	Gap Between	What is the gap in days between "José Luis Soro became the leader of Chunta Aragonesista" to "Maru Díaz became the leader of Podemos–Green Alliance in Aragon." and "Tomás Guitarte became the leader of Teruel Existe" to "Alberto Izquierdo became the leader of the Aragonese Party.", in the context of "2023 Aragonese regional election"?	1522
	Inclusion	Which event's time span includes the other: "June 2009 Extra Session of the 99th Wisconsin Legislature ended" to "December 2009 Special Session of the 99th Wisconsin Legislature began", or "Election for the 99th Wisconsin Legislature was held" to "The term of the 99th Wisconsin Legislature ended.", in the context of "99th Wisconsin Legislature"?	Election for the 99th Wisconsin Legislature was held to The term of the 99th Wisconsin Legislature ended.
	Order Span End	Which event span ended last: "The album '100 Reasons to Live' by Gareth Emery was released" to "The single 'Far From Home' by Gareth Emery feat. Gavrielle was released", or "The single 'Hands' by Gareth Emery &Alastor feat. London Thor was released" to "The single 'Save Me' by Gareth Emery was released.", in the context of "100 Reasons to Live"?	"The single 'Hands' by Gareth Emery &Alastor feat. London Thor was released" to "The single 'Save Me' by Gareth Emery was released."
	Order Span Start	Which event span started earlier: "The album '100 Reasons to Live' by Gareth Emery was released" to "The single 'Far From Home' by Gareth Emery feat. Gavrielle was released", or "The single 'Hands' by Gareth Emery &Alastor feat. London Thor was released" to "The single 'Save Me' by Gareth Emery was released.", in the context of "100 Reasons to Live"?	"The single 'Hands' by Gareth Emery &Alastor feat. London Thor was released" to "The single 'Save Me' by Gareth Emery was released."
	Overlap	How many days do "The single 'Hands' by Gareth Emery &Alastor feat. London Thor was released" to "The single 'Far From Home' by Gareth Emery feat. Gavrielle was released.", and "The album '100 Reasons to Live' by Gareth Emery was released" to "The single 'Save Me' by Gareth Emery was released." overlap, in the context of "100 Reasons to Live"?	29
CountQA		The following is a list of historical events. Each line includes a description and the title of the article it comes from. (1) Event about 'Sean Kingston (album)': The single "Beautiful Girls" by Sean Kingston was released. (2) Event about 'Week End (X Japan song)': The song "Week End" by X Japan was released. (3) Event about 'Lucy McEvoy': Lucy McEvoy was nominated for the AFL Women's Rising Star award. (4) Event about 'BL 5.4-inch howitzer': The Ordnance BL 5.4-inch howitzer was used in the Second Boer War. (5) Event about 'Holten Castenschiold': Holten Castenschiold began his term as the 6th President of the Danish Olympic Committee. (6) Event about 'William L. Baird': William Lewis Baird began his term as the 19th Mayor of Lynn, Massachusetts. (7) Event about 'Jung Yeon-kyung': Jung Yeon-kyung was born. Provide the count of the number of events that occurred during the 19th century.	2
SequenceQA		Sort the events by the time they took place. Each event is accompanied by a description and the title of the article it comes from. Return the correct chronological order by listing the event numbers, like (2) (1) (3). (1) Event about 'Paaliaq': Paaliaq was discovered. (2) Event about 'Bob Chakales': Bob Chakales was born. (3) Event about 'Dark Valley': The film "Dark Valley" was released.	(2) (3) (1)
RecurrenceQA		The following event includes the article title and event description. Event about '2008 Iranian legislative election': The second round of the 2008 Iranian legislative election was held. Identify the year when the previous edition of the event took place. Provide the year as your answer.	2004

Table 1: Examples of English Questions from our HistoryBankQA dataset. More examples are in Tables 11, 12 and 13 in Appendix G.

despite their value for modeling factual consistency, event progression, and causal relations over long timelines. Such structured knowledge is vital for tasks like timeline summarization, temporal QA, and evaluating LLMs' temporal reasoning. However, building high-quality, temporally anchored historical event datasets remains challenging, requiring synthesis of semi-structured sources, temporal disambiguation, and aligning event granularity across languages.

To address this, we propose HistoryBank, a large-scale multilingual event database (DB) built from Wikipedia's event-centric pages, including *On This Day* listings[2] and infoboxes. The database spans ten languages: English (en), Bengali (bn), German (de), French (fr), Indonesian (id), Hindi (hi), Italian (it), Portuguese (pt), Russian (ru), and Spanish (es). It covers diverse domains such as politics, culture, science, and sports. For English alone, it includes ∼8.2M events, making it the largest publicly available multilingual temporal event resource.

Towards the second challenge, recent benchmarks like TRAM (Wang and Zhao, 2024), TG-LLM (Xiong et al., 2024), TempReason (Tan et al., 2023), and ChronoSense (Islakoglu and Kalo, 2025) assess LLMs' temporal reasoning but mostly use synthetic setups, abstract events, or daily contexts rather than historically grounded contexts. Our work instead targets reasoning over curated, encyclopedic historical events, enabling joint evaluation of memorization (factual recall) and reason-

[2]For example, for English, this page links to month-wise pages which link to a page per day of the year: https://en.wikipedia.org/wiki/Category:Days_of_the_year.

ing (ordering, comparison, disambiguation) in a realistic setting. This dual challenge better reflects real-world expectations, where users need models to reason over temporally structured factual knowledge, not just toy examples or synthetic timelines.

Further, our multilingual design spans 10 typologically diverse languages, allowing multilingual analysis of temporal reasoning, an underexplored area. Our motivation for a multilingual benchmark is that temporal reasoning failures often arise from linguistic and cultural variation in how time is expressed. Temporal expressions, date formats, aspect, tense, and event-ordering cues differ significantly across languages (e.g., implicit vs. explicit temporal connectives, script differences, relative vs. absolute tense usage). Evaluating only in English provides an incomplete picture and risks overstating model capability, because models can exploit English-specific cues. A multilingual benchmark exposes whether failures are due to limitations in temporal reasoning or merely in English-specific heuristics, thus providing a clearer picture of the underlying reasoning ability. Many temporal constructs are language-specific and can introduce reasoning challenges that do not appear in English.

Using our event corpus, we build HistoryBankQA, a suite of six temporal QA tasks: (i) explicit date retrieval (FactQA), (ii) event sequencing and ordering (SequenceQA), (iii) interval computation (DurationQA), (iv) relative time expression resolution (RelationQA), (v) event counting within periods like a century or between two events/dates (CountQA), and (vi) identifying previous or next occurrences of a recurring event (RecurrenceQA). Each dataset systematically probes model performance across languages and event types. Table 1 shows an example per question type; more appear in Tables 11, 12 and 13 in Appendix G.

We make the following main contributions: (1) **Dataset**: We introduce HistoryBank, the largest publicly available multilingual historical event database, with 10M+ events across 10 languages, built from Wikipedia's On This Day pages and event-centric infoboxes. (2) **Benchmark**: We design a comprehensive suite of six multilingual temporal QA tasks grounded in real historical events for robust multilingual and cross-domain evaluation of temporal understanding in LLMs. (3) **Initial Findings**: We report zero-shot and initial RAG-based results for four small language models and GPT-4o, highlighting strengths and limitations in temporally grounded factual reasoning, and release code, event database, and QA benchmarks[1].

Language	#Articles	#Infoboxes	#Events	#Events per Infobox	#Events per Article
English	4,260,757	4,729,733	8,201,400	1.73	1.92
Bengali	47,350	52,166	78,181	1.50	1.65
German	998,020	1,085,679	1,519,208	1.40	1.52
French	113,489	199,910	112,207	0.56	0.99
Indonesian	314,856	346,912	66,096	1.90	2.10
Hindi	50,946	59,885	162,717	2.71	3.19
Italian	6,548	6,766	19,473	2.87	2.97
Portuguese	10,774	12,846	10,201	0.79	0.95
Russian	5,756	6,915	7,979	1.15	1.39
Spanish	8,688	19,847	23,536	1.18	2.71

Table 2: Language-wise HistoryBank Statistics

2 HistoryBank: A Large-scale Multilingual DB of Historical Events

We build a multilingual event database from Wikipedia by extracting temporally grounded facts from infoboxes which are semi-structured templates summarizing salient entity attributes via key-value pairs (e.g., `born = January 9, 1913`). Many of these values are temporally grounded, such as birth dates, foundation years, or historical milestones, which can be interpreted as implicit events. These entries are transformed into natural language events (e.g., "Richard Nixon was born on January 9, 1913."). To extract such events, we prompt GPT4o for each infobox across the 10 languages to: (1) detect date-bearing key-value pairs, (2) parse dates, (3) generate concise event descriptions, (4) retain source keys, (5) annotate relevant countries, and (6) assign country-specific polarity (positive/negative/neutral). Full prompt details appear in Appendix C.

Infoboxes are prioritized over free-text content for structural consistency, enabling reliable parsing. To address generic event descriptions (e.g., "The individual played for the Oakland Raiders"), we prepend the article title to each event description. As shown in Table 2, English Wikipedia yields 8M+ events from 4.7M infoboxes (avg. 1.73 per infobox); other languages contribute varying densities, forming a diverse multilingual corpus. Significant variation in events per infobox and events per article underscores the differing conventions across languages in how infoboxes are employed. Event distribution (Fig. 1) shows steady growth over time, accelerating post-1960, with 11,687 events dated before 0 (BC). Appendix B provides further analysis, including time-wise distributions across languages (Fig. 3), infobox-type (or entity type) distributions (Fig. 4 for English, Table 8 for other languages), top 50 countries (Table 9), and polarity

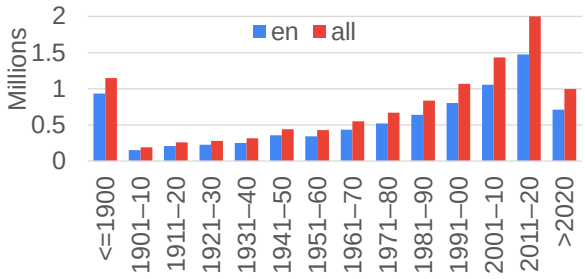

Figure 1: Time-wise Distribution of Events (English vs all languages) in HistoryBank.

distribution (Fig. 5).

We also included an additional dataset sourced from Wikipedia's On This Day pages. To collect this data, we automatically downloaded and organized historical event information (about births, deaths and generic events) from Wikipedia's day-specific pages. The extracted information was standardized by associating each entry with a list of relevant countries and a polarity label. Table 10 in Appendix B shows data statistics per language.

3 HistoryBankQA: Temporal QA Benchmark

One effective way to evaluate LLMs' temporal reasoning is via temporal QA. Our multilingual suite, HistoryBankQA, comprises six QA tasks (*FactQA*, *DurationQA*, *RelationQA*, *SequenceQA*, *CountQA*, *RecurrenceQA*), each targeting a specific temporal reasoning capability, All of these task datasets are derived from our HistoryBank dataset introduced in Section 2 and share a consistent schema across ten languages. These tasks target complementary temporal capabilities and capture coarse-grained temporal understanding, addressing gaps in existing evaluations. Table 3 reports per-task, per-language example counts.

Our task design is grounded in well-established components of temporal reasoning that are essential for understanding event structure. Specifically, the benchmark evaluates key dimensions such as duration (DurationQA), ordering (SequenceQA), recurrence (RecurrenceQA), and typical or canonical time (FactQA). These categories reflect fundamental operations required for temporal understanding. Some of our tasks intentionally combine multiple reasoning skills to reflect more realistic scenarios. For example, CountQA requires models to integrate typical-time knowledge with counting-based temporal inference, thereby testing compositional or combined reasoning. Such formulations align with prior work on temporal and event-based reasoning, including Zhou et al. (Zhou et al., 2019) and Wang and Zhao (Wang and Zhao, 2024), which similarly decompose temporal understanding into these core components.

FactQA: Factual Date Identification. FactQA evaluates a model's ability to recall the date on which a specific event occurred, given a natural language question. For each extracted event in HistoryBank, we use its associated title, event description, and temporal key to construct a self-contained context. Using this context, we prompt an instruction-tuned generative LM (RecurrentGemma-2B-IT (Botev et al., 2024) for English and Gemma-3B-IT (Team et al., 2025) for other languages) to generate a natural question that asks for the event's timing. This pipeline is applied uniformly across all ten languages, ensuring linguistic diversity and contextual fidelity. The prompt is listed in Appendix D.

DurationQA: Temporal Span Reasoning. DurationQA focuses on determining the length of a real-world event, requiring inference of the elapsed time between its start and end points. We collect events within an article, sort them chronologically, form start–end event pairs, and filter valid pairs with a LM classifier (Flan-T5-XL (Chung et al., 2024) for English; Gemma-2B-IT for other langs). Filtered pairs are sent to the generator (RecurrentGemma-2B-IT / Gemma-3B-IT) to create questions about time elapsed between the two events. Prompts are in Appendix E.

RelationQA: Temporal Span Comparison. RelationQA evaluates comparative reasoning over pairs of event spans (e.g., which lasted longer, whether they overlapped, which began earlier). For each article with $\geq$2 spans, we generate all unique unordered span pairs, compute duration, overlap, inclusion, gap, and relative order, and exclude spans with identical endpoints. Each pair is annotated with up to seven question types: (1) *Compare Duration* (assesses relative length), (2) *Duration Difference* (in days), (3) *Overlap* (in days), (4) *Gap Between* (number of days separating the two spans), (5) *Inclusion* (whether one span is entirely contained within the other), (6) *Order Start* (which span began earlier), (7) *Order End* (which span concluded later). Questions are generated using templates anchored in the event descriptions and entity context; answers are derived automatically from the dates.

SequenceQA: Chronological Ordering Of Events. SequenceQA is a QA task focused on the

Task		English	Bengali	German	French	Indonesian	Hindi	Italian	Portuguese	Russian	Spanish
FactQA		4413790	38479	765258	70774	468121	86814	13341	4842	3888	19877
DurationQA		1731411	5549	30859	2250	47948	7328	312	123	482	308
RelationQA	Compare Duration	1243222	5789	5460	8583	28677	7178	46	46	20	64
	Duration Diff	1449966	5782	5457	8532	28506	7098	46	46	20	64
	Gap Between	873727	684	760	1611	4344	1210	6	7	5	16
	Inclusion	839775	1219	583	1314	7193	1449	1	4	9	9
	Order Span End	2120530	2936	1700	3766	13902	3623	9	11	12	32
	Order Span Start	2120530	2936	1700	3766	13902	3623	9	11	12	32
	Overlap	1742192	5124	4707	6924	24718	6058	40	42	15	48
CountQA		1500000	14000	270000	20000	12000	29000	3500	1850	1450	4200
SequenceQA		2100000	19545	379802	28052	16524	40679	4868	2550	1995	5884
RecurrenceQA		42620	172	0	0	2243	258	0	0	0	0

Table 3: Number of examples per temporal reasoning task and language in our proposed HistoryBankQA benchmark.

correct sequencing of historical events. This task captures three primary temporal reasoning sub-task types, each framed as a distinct QA format. (1) Sequence Arrangement (Freeform Order Prediction): Models are given a set of unordered events and asked to arrange them in chronological order. Task is described using this example sentence: "Arrange the following events in the order they occurred." (2) Multiple Choice Question (MCQ): Models are presented with multiple candidate sequences and must choose the one that is temporally accurate. Task is described using this example sentence: "Which of the following sequences places the events in the right chronological order?" (3) Order Verification (True/False): A single sequence of events is provided, and models must determine whether the sequence is chronologically correct. Task is described using this example sentence: "Is this the correct historical order of the events?"

To avoid any structural biases and encourage generalization, each task description sentence was diversified using 10 different natural language syntactic variant question templates.

CountQA: Counting Events. The CountQA dataset consists of question-answer pairs that require models to reason about the number of historical events within specific temporal bounds. For each question, we sample 4 to 7 events, and design three primary types of counting questions as follows. (1) Count Between Events: Given a sequence of events (not necessarily in chronological order), the question asks how many events fall chronologically between two specific events in the list. Example: "How many events occurred between event (2) and event (5)?" (2) Count Between Years: A time range is specified using two randomly sampled years ending in 00 (e.g., 1200-1400), and the model must count how many events from the list fall within that range. Example: "How many of the events occurred between the years 1500 and 1700?" (3) Count by Century: The model is asked to count how many events occurred in a given century (e.g., 19th century). Centuries are sampled according to their observed frequency distribution in the data to ensure realistic coverage. Example: "How many events below occurred in the 20th century?"

Each question is generated using natural language paraphrase templates to introduce diversity in phrasing while maintaining clarity.

RecurrenceQA: Identifying Previous Occurrences Of Recurrent Events. To systematically evaluate temporal reasoning over recurring events (such as elections, sports leagues, or international tournaments), we construct a dataset consisting of questions that require identifying the year when a prior occurrence of a given event occurred. Candidate events are first identified by selecting entries whose metadata indicates a temporal relationship to an earlier occurrence. For example, if an event record (e.g., https://en.wikipedia.org/wiki/2011_Cricket_World_Cup) explicitly refers to a previous year or season (e.g., https://en.wikipedia.org/wiki/2007_Cricket_World_Cup), we interpret this reference as evidence of recurrence. To validate these relationships, we retrieve the corresponding current version of the event within the same article group and verify that the annotated date fields are well-formed and temporally consistent; the year of the current event must exceed the year of the referenced prior event. This validation step ensures that the resulting questions are anchored in correct chronological sequences.

The question asks the year in which the earlier version occurred with a natural language question. To increase linguistic diversity, the prompt is randomly sampled from a curated set of paraphrase templates, which rephrase the temporal query in varied ways. All questions are phrased to elicit a precise, single-year answer. This approach pro-

duces a collection of temporally grounded QA pairs designed to evaluate whether models can recall/ infer timing of recurring events. By explicitly requiring models to resolve cross-references between related historical records, the task provides a targeted test of temporal retrieval and factual consistency.

4 Experiments

To assess the quality and difficulty of our temporal reasoning benchmarks, we randomly sampled up to 10,000 test examples from each task per language. For languages or benchmarks with fewer examples, we used the entire available set. This leads to a test data with 535843 samples across all languages. We evaluated these samples using a suite of multilingual LLMs with diverse architectural and parameter configurations: GPT4o-1120 (OpenAI et al., 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), Gemma-2-9B-IT (Riviere et al., 2024), Llama-3-8B-Instruct (Grattafiori et al., 2024), and Qwen3-8B (Yang et al., 2025). This selection includes both closed and open models, enabling a broad assessment of capabilities across scales and training paradigms.

Our benchmark tasks are designed to probe two complementary aspects of temporal reasoning: (1) Historical factual knowledge, e.g., recognizing when events occurred, even without temporal cues. (2) Temporal inference and reasoning over distant or multi-event sequences, requiring models to understand event ordering and long-range temporal relationships.

Importantly, we evaluate these models in a zero-shot setting without fine-tuning. While fine-tuning can improve reasoning patterns, it cannot substitute for the breadth and accuracy of factual historical knowledge, which must be present in the model's pretraining. This setup ensures that our benchmark serves as a faithful diagnostic of intrinsic temporal reasoning ability, rather than post-hoc adaptation. Prompts used are mentioned in Appendix F.

We also experiment with a retrieval augmented generation (RAG) based setup. We index all infoboxes and create a FAISS index per language. We used the intfloat/multilingual-e5-small embeddings for bn, fr, de, hi, id, ru and es. We used all-MiniLM-L6-v2 embeddings for en. For Italian, we used dbmdz/bert-base-italian-xxl-cased embeddings. Lastly, for Portuguese, we used neuralmind/bert-base-portuguese-cased embeddings. We use 10 retrievals for RAG

	Task	en	bn	de	fr	id	hi	it	pt	ru	es
	FactQA	48.8	48.9	43.3	47.2	60.3	52.5	36.6	65.9	58.1	58.9
	DurationQA	31.8	80.5	88.6	79.1	89.3	81.5	88.9	87.5	90.6	82.6
RelationQA	Compare Dur.	73.2	55.7	77.7	61.5	77.1	73.7	82.6	73.9	-	57.8
RelationQA	Duration Diff	13.8	6.5	30.2	9.4	32.7	24.3	19.6	43.5	-	14.8
RelationQA	Gap Between	13.1	6.7	25.5	6.8	35.6	24.3	-	-	-	-
RelationQA	Inclusion	70.7	8.3	80.6	38.8	86.3	58.5	-	-	-	-
RelationQA	Order Span End	60.0	58.2	58.6	54.8	84.7	71.5	-	-	-	37.5
RelationQA	Order Span Start	84.1	79.0	88.8	72.3	91.3	84.8	-	-	-	84.4
RelationQA	Overlap	10.8	6.0	29.5	10.0	19.2	21.0	20.0	41.7	-	16.7
	CountQA	38.3	31.9	38.2	38.4	38.3	37.8	38.2	45.6	40.8	46.1
	SequenceQA	54.1	52.2	51.8	50.0	54.9	54.5	55.8	64.7	63.8	55.5
	RecurrenceQA	82.5	89.0	-	-	85.8	95.2	-	-	-	-

Table 4: GPT4o performance across languages on temporal reasoning tasks in our proposed HistoryBankQA benchmark. We don't report results for cells with test sample size < 30.

	Task	en	bn	de	fr	id	hi	it	pt	ru	es
CountQA	Century	48.6	47.4	49.2	49.1	52.6	53.6	46.0	54.3	54.1	61.3
CountQA	Bet_events	20.7	12.5	19.2	19.1	18.6	14.7	18.3	20.4	20.2	16.6
CountQA	Bet_dates	45.7	35.6	45.9	46.9	44.9	44.4	51.1	62.4	47.2	58.8
SeqQA	Verify	55.6	51.1	57.7	53.0	53.2	52.7	57.6	61.9	62.9	54.7
SeqQA	MCQ	65.9	65.0	61.4	59.8	67.6	66.8	68.8	75.9	76.4	67.1
SeqQA	Arrange	40.3	40.1	36.7	37.1	44.4	43.6	40.7	56.7	52.3	44.9

Table 5: Detailed performance of GPT4o across languages for CountQA and SequenceQA tasks in our proposed HistoryBankQA benchmark.

5 Results

We report exact match accuracy for all the 6 tasks across languages. Results for GPT4o are reported in Table 4. Similarly, results for Mistral-7B-Instruct, LLaMA-3-8B-Instruct, Gemma-2-9B-IT, and Qwen3-8B are reported in Table 6. We also present results using our RAG experiments in Tables 14 and 15 in Appendix J. Lastly, we discuss human evaluation results and present a qualitative error analysis.

Performance by Task Type and Language. We plot the accuracy for each model by task type in Fig. 2. Fig. 2 (right) specifically focuses on subtasks within RelationQA. We observe significant variation in model performance across different temporal reasoning tasks. The lowest accuracy across all models is consistently seen in the *Gap Between*, *Overlap*, and *Duration Difference* tasks within the RelationQA benchmark. These tasks demand a multi-step reasoning process that combines temporal information extraction with numerical inference. Specifically, models must correctly (1) identify the start and end dates of two separate event spans, (2) parse these dates into structured formats, and (3) compute the corresponding difference (e.g., number of overlapping days, total gap

Mistral-7B-Instruct		Task	en	bn	de	fr	id	hi	it	pt	ru	es
		FactQA	26	8	22	20	21	10	21	25	23	26
		DurationQA	24	65	68	62	67	63	66	65	66	64
	RelationQA	Compare Duration	36	12	40	33	35	30	39	37	-	34
	RelationQA	Duration Diff	10	1	14	8	13	8	11	14	-	9
	RelationQA	Gap Between	10	1	10	5	12	8	-	-	-	-
	RelationQA	Inclusion	40	5	50	30	58	45	-	-	-	-
	RelationQA	Order Span End	45	40	48	43	60	50	-	-	-	35
	RelationQA	Order Span Start	60	55	70	60	78	68	-	-	-	68
	RelationQA	Overlap	8	2	10	6	8	7	7	10	-	8
		CountQA	29	8	28	24	18	2	27	25	27	26
		SequenceQA	17	3	17	11	4	10	14	10	16	14
		RecurrenceQA	38	2	-	-	43	2	-	-	-	-

LLaMA-3-8B-Instruct		Task	en	bn	de	fr	id	hi	it	pt	ru	es
		FactQA	12	10	8	9	11	13	10	14	12	13
		DurationQA	10	30	35	30	35	32	33	32	34	32
	RelationQA	Compare Duration	25	18	27	20	26	28	30	27	-	21
	RelationQA	Duration Diff	5	2	6	3	7	6	4	6	-	4
	RelationQA	Gap Between	5	1	4	2	7	5	-	-	-	-
	RelationQA	Inclusion	25	3	30	18	32	26	-	-	-	-
	RelationQA	Order Span End	30	28	32	27	36	35	-	-	-	24
	RelationQA	Order Span Start	45	40	50	42	53	50	-	-	-	50
	RelationQA	Overlap	4	1	5	3	4	5	4	7	-	4
		CountQA	5	8	7	9	5	5	9	10	11	8
		SequenceQA	10	10	14	14	17	17	14	17	15	10
		RecurrenceQA	65	49	-	-	52	87	-	-	-	-

Gemma-2-9B-IT		Task	en	bn	de	fr	id	hi	it	pt	ru	es
		FactQA	41	30	38	39	40	36	35	42	41	40
		DurationQA	60	75	79	76	81	73	78	79	82	78
	RelationQA	Compare Duration	62	45	65	59	64	61	67	64	-	63
	RelationQA	Duration Diff	20	8	22	18	24	22	20	27	-	21
	RelationQA	Gap Between	18	7	20	17	25	21	-	-	-	-
	RelationQA	Inclusion	60	15	70	45	78	62	-	-	-	-
	RelationQA	Order Span End	65	55	70	60	80	68	-	-	-	68
	RelationQA	Order Span Start	78	70	85	75	90	82	-	-	-	85
	RelationQA	Overlap	15	5	18	12	16	14	13	20	-	15
		CountQA	28	21	28	27	26	24	32	30	29	29
		SequenceQA	35	26	34	34	31	34	31	39	37	32
		RecurrenceQA	77	71	-	-	80	88	-	-	-	-

Qwen3-8B		Task	en	bn	de	fr	id	hi	it	pt	ru	es
		FactQA	35	15	31	32	33	24	30	36	35	37
		DurationQA	55	70	72	71	74	68	73	74	76	72
	RelationQA	Compare Duration	55	30	50	48	51	45	53	52	-	54
	RelationQA	Duration Diff	16	4	18	15	19	16	17	23	-	20
	RelationQA	Gap Between	15	3	15	12	18	14	-	-	-	-
	RelationQA	Inclusion	55	10	60	42	70	50	-	-	-	-
	RelationQA	Order Span End	58	47	62	55	73	60	-	-	-	64
	RelationQA	Order Span Start	73	65	80	70	88	75	-	-	-	80
	RelationQA	Overlap	12	3	14	10	13	11	11	16	-	13
		CountQA	27	6	22	26	20	10	20	21	26	27
		SequenceQA	28	9	22	23	22	11	25	22	29	25
		RecurrenceQA	37	22	-	-	22	35	-	-	-	-

Table 6: Mistral-7B-Instruct (top left), LLaMA-3-8B-Instruct (top right), Gemma-2-9B-IT (bottom left), Qwen3-8B (bottom right) performance across languages on temporal reasoning tasks in our proposed HistoryBankQA benchmark. We don't report results for cells with test sample size less than 30.

between spans, or absolute duration differences).

This multi-hop requirement introduces substantial error propagation. Even a small mistake in identifying one of the event dates can cascade into an incorrect final answer. By contrast, tasks like *Order Span Start* show comparatively better performance, possibly because determining which event started earlier can sometimes be resolved using partial temporal cues, even if both spans are not fully parsed.

Sequence ordering tasks also demonstrate moderate performance, suggesting that models are more comfortable with qualitative temporal relationships (e.g., before/after) than quantitative reasoning over time intervals. Overall, these results highlight the challenges models face when asked to internalize structured time and perform arithmetic or set-based reasoning over date spans.

Table 5 lists the sub-task results for CountQA and SequenceQA for GPT4o. Within the CountQA tasks, the *Between_events* setting shows the lowest performance. This is expected, as the task implicitly combines three subtasks: (i) identifying the event dates, (ii) sorting the events in temporal order, and (iii) reasoning about how many fall between two given events. Each of these steps introduces potential sources of error, leading to compounding difficulty. In contrast, the *Between_dates* variant is relatively easier, since it only requires finding the event dates and counting how many lie within a fixed range, without the added complexity of event-to-event comparisons. The *Century* type is somewhat easier still, as it reduces to recognizing the century of given events, which often boils down to identifying whether an event happened in the previous edition or within a well-bounded historical period.

For the SequenceQA tasks, we also see a natural ordering of difficulty. The *MCQ* setting achieves the highest performance, as the model only needs to identify the correct sequence from a small number of options, which effectively turns the task into a recognition problem. The *Verify* variant is more challenging, requiring the model to judge the correctness of an event sequence in a binary True/False format. Finally, the *Arrange* type is the most difficult, since it demands constructing the chronological order from scratch by reasoning over multiple events, which amplifies the complexity of temporal ordering. Overall, these trends highlight that performance is strongly correlated with the number of reasoning steps involved: tasks requiring multi-step temporal reasoning (e.g., Between_events, Arrange) consistently exhibit lower scores compared to those that reduce to direct recognition or simpler range checks.

We plot the accuracy for each model by language in Fig. 2 (left). Interestingly, despite being relatively low-resource in terms of Wikipedia coverage compared to English or German, Hindi and Indonesian consistently yield the highest average model performance across all benchmark

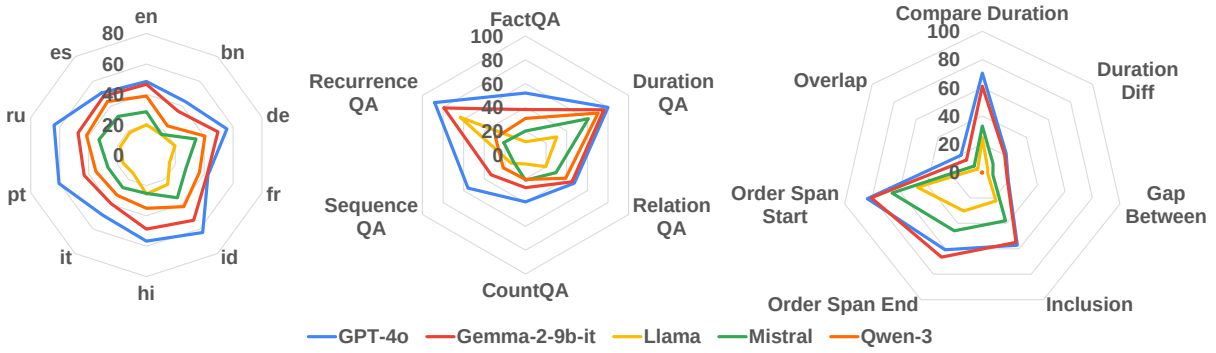

Figure 2: Performance by (left) Language (middle) Question Type (right) Question Type for RelationQA sub-tasks

tasks. These findings underscore that model perf is not solely a function of language scale but also of how temporally structured information is represented in that language's Wikipedia and how well the model captures its grammatical and formatting conventions.

Performance using various LLMs. When compared against GPT4o, the four smaller models: LLaMA-3-8B-Instruct, Mistral-7B-Instruct, Gemma-2-9B-IT, and Qwen-3-8B show much lower accuracies across almost all tasks and languages. Nonetheless, clear differences emerge in their relative capabilities.

Among the smaller models, Gemma-2-9B-IT consistently achieves the highest accuracies across most tasks, often approaching GPT4o performance on structured reasoning problems such as DurationQA, Compare Duration, and Order Span Start. Qwen-3-8B ranks second, showing robust but slightly lower performance than Gemma, particularly on arithmetic-heavy tasks. Mistral-7B-Instruct demonstrates moderate accuracy, performing best in qualitative ordering tasks (e.g., Order Span Start/End) but faltering in numerical reasoning. LLaMA-3-8B-Instruct is the weakest overall, with very low performance on tasks that require multi-step reasoning (Duration Difference, Gap Between, and Overlap). This ranking establishes a clear hierarchy: Gemma $>$ Qwen $>$ Mistral $>$ LLaMA.

Unlike GPT-4o, which maintains relatively stable multilingual performance (particularly in high-resource languages such as Spanish, Portuguese, and Russian) smaller models show pronounced language-specific degradation. LLaMA performs especially poorly in Bengali and Hindi across most tasks, often dropping below 10% accuracy on harder tasks like Gap Between and Overlap. Mistral exhibits high cross-lingual variance, with sharp collapses in Bengali and Hindi but partial recovery in Romance languages. Gemma and Qwen are comparatively more balanced, yet still underperform in low-resource languages relative to GPT-4o.

Task-level Trends. The task-level trends of smaller models broadly mirror those observed for GPT4o: (1) *Gap Between*, *Overlap*, and *Duration Difference* remain the hardest tasks, with all smaller models performing even worse than GPT4o (often below 20). This highlights that fine-grained interval arithmetic is a general weakness, exacerbated in smaller models. (2) Ordering tasks (*Order Span Start/End*) are comparatively easier. Gemma in particular achieves > 80 in several languages, nearly matching GPT4o, while LLaMA lags significantly. (3) *DurationQA* shows the largest gap between GPT4o and the smaller models: while GPT4o averages ~ 80, Gemma and Qwen reach only 70–78, with Mistral and LLaMA trailing far behind. (4) *Inclusion* is another area where Gemma and Qwen approach GPT4o, with accuracies > 70 in many Romance languages, while LLaMA and Mistral remain far weaker. (5) *CountQA* and *SequenceQA* are uniformly low across all models, but GPT4o still outperforms the smaller models by a wide margin.

Despite these differences, the overall task difficulty ranking remains consistent with GPT4o: ordering and qualitative reasoning are easier, while numerical reasoning over temporal spans remains the most challenging.

Performance using RAG. Table 14 in Appendix J presents the RAG results with GPT-4o as the generation model. Table 15 in Appendix J reports corresponding results using Mistral-7B-Instruct, LLaMA-3-8B-Instruct, Gemma-2-9B-IT, and

Qwen3-8B. We do not observe any gains compared to the results shown in Tables 4 and 6. Perhaps, this is because (1) our questions depend on events extracted from infoboxes and events do not have large verbose content to contextualize for RAG experiments, (2) identifying the correct event spans requires near-perfect retrieval accuracy over millions of events, and even small retrieval errors propagate directly into incorrect answers. Manual inspection of retrieved documents for a subset of queries reveals generally low retrieval quality, which likely limits downstream performance. Overall, these results indicate that temporal reasoning remains a challenging setting for retrieval-augmented generation and suggest that task-specific fine-tuning could be more effective than relying on simple RAG-based approaches alone.

Human Evaluation. We conduct a human evaluation on 100 randomly sampled English events to assess generation quality. Annotators evaluate each event along three dimensions: Faithfulness, Event Interpretability, and Event Interestingness. Faithfulness measures whether the generated event description accurately reflect the infobox information. Event Interpretability measures whether the event understandable as a standalone historical event. Event Interestingness measures whether the discovered event meaningfully corresponds to a real-world historical event. We provide detailed human annotation guidelines in Appendix I.

Results show that 97% of events are fully faithful to the source infobox information, with the remaining 3% exhibiting only minor issues and no unfaithful generations, indicating that the model reliably grounds event descriptions in source facts. In terms of interpretability, 88% of events are rated as clear standalone events, while 6% are partially clear, primarily due to descriptions that rely on the Wikipedia article title for full context (e.g., "The television show last aired"), motivating our recommendation to present the article title alongside the event description. Finally, 84% of events are judged as interesting, with the remaining 16% deemed uninteresting, typically corresponding to low-salience infobox entries such as administrative mentions, routine statistical updates, or census records (e.g., population counts). Overall, the results indicate high factual reliability of the generated events, with remaining limitations largely driven by infobox entries that are weakly eventful rather than by generation errors.

Error Analysis. We analyze incorrect predictions on the English HistoryBankQA test set using GPT-4o to identify recurring sources of error in temporal question answering. Manual inspection reveals that errors arise at different stages of the question-answering process and are not confined to a single task type or dataset. In particular, failures are more likely to occur when tasks require multiple latent reasoning steps: such as identifying or inferring event dates, applying temporal boundaries, ordering events, and aggregating counts, though such errors are also observed in simpler settings.

Across examples, we identify three recurring failure modes. First, models sometimes lack or misrepresent the required temporal knowledge, resulting in incorrect event dates or refusals that cite insufficient information even when inference from context or world knowledge is expected. Second, in other cases the knowledge is correct but the reasoning is flawed, manifesting as erroneous temporal comparisons, inconsistent boundary handling (e.g., century definitions or inclusive vs. exclusive ranges), or misordered events despite correct dates. Third, we observe reasoning–answer mismatches, where the knowledge and intermediate reasoning are correct, but the final answer contradicts the stated reasoning.

6 Conclusion

We introduced HistoryBank, a large-scale multilingual dataset of historical events automatically extracted from Wikipedia timelines and article infoboxes. We showed that both the scale and structure of extracted events vary widely across languages, with English contributing the largest share. Using this event database, we built HistoryBankQA, a comprehensive benchmark of temporal reasoning tasks that evaluate models' ability to understand and reason about time in real-world historical contexts. We further established baseline results with several state-of-the-art multilingual language models. Our results reveal substantial performance variation across both tasks and languages, highlighting the linguistic and temporal reasoning challenges inherent in multilingual historical data.

Limitations

While our work presents a significant advancement in multilingual temporal reasoning over historical

events, several limitations remain that warrant future exploration:

- **Event Coverage Bias:** Our event database is constructed primarily from Wikipedia infoboxes and On This Day pages, which tend to emphasize Western-centric, encyclopedic, and notable events. This introduces a coverage bias, potentially underrepresenting events from marginalized regions, indigenous cultures, and non-mainstream historical narratives.
- **Synthetic Question Generation:** The QA datasets rely on automatically generated questions using instruction-tuned models. While this enables scalability, it may introduce artifacts such as unnatural phrasing, limited diversity, or subtle biases in question formulation. Human-authored questions could improve linguistic richness and realism.
- **Limited Evaluation Scope:** Our baseline evaluations focus on zero-shot setting using a small set of models. We do not explore few-shot learning, retrieval augmented generation (RAG), or more advanced prompting strategies, which may yield better performance and deeper insights into model capabilities.
- **Temporal Reasoning Beyond Factual Recall:** While our benchmark includes tasks that go beyond simple date retrieval, many questions still rely on factual recall rather than complex temporal inference involving causality, counterfactuals, or narrative coherence. Expanding the benchmark to include such dimensions would further enrich the evaluation landscape.
- **Lack of Human Evaluation:** We rely on automatic metrics and answer keys derived from structured data. Human evaluation of model responses, especially for open-ended or comparative tasks, could provide more nuanced insights into reasoning quality and answer plausibility.

7 Ethical Considerations

This work involves the construction of a large-scale multilingual dataset of historical events and the development of temporal question answering benchmarks. All data used in this study are derived from publicly available sources, primarily Wikipedia, which is licensed under Creative Commons Attribution-ShareAlike (CC BY-SA). We respect the licensing terms and ensure that our dataset does not include any proprietary or personally identifiable information.

The automatic extraction and generation processes rely on large language models (LLMs), which may introduce biases or inaccuracies. We acknowledge that Wikipedia content itself may reflect cultural, geographic, and editorial biases, which can propagate into our dataset. We have taken steps to mitigate these effects by including diverse languages and domains, but further work is needed to ensure equitable representation across cultures and historical perspectives.

Our benchmark tasks are designed for research purposes only and should not be used in high-stakes decision-making without human oversight. We caution against deploying temporal reasoning models trained on this dataset in sensitive applications such as historical education, policy analysis, or legal contexts without rigorous validation.

Finally, we commit to releasing our dataset and benchmarks under an open license to support transparency, reproducibility, and community-driven improvements. We encourage responsible use and welcome feedback to improve the fairness, inclusivity, and accuracy of our resources.

References

Abdelrahman Abdallah, Mohammed Ali, Muhammad Abdul-Mageed, and Adam Jatowt. 2026. Tempo: A realistic multi-domain benchmark for temporal reasoning-intensive retrieval. *arXiv preprint arXiv:2601.09523*.

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pages 722–735. Springer.

Claire Barale, Leslie Barrett, Vikram Sunil Bajaj, and Michael Rovatsos. 2025. Lextime: A benchmark for temporal ordering of legal events. *arXiv preprint arXiv:2506.04041*.

Nadav Borenstein, Natalia da Silva Perez, and Isabelle Augenstein. 2023. Multilingual event extraction from historical newspaper adverts. *arXiv preprint arXiv:2305.10928*.

Aleksandar Botev, Soham De, Samuel L Smith, Anushan Fernando, George-Cristian Muraru, Ruba Haroun, Leonard Berrada, Razvan Pascanu, Pier Giuseppe Sessa, Robert Dadashi, and 1 others. 2024. Recurrentgemma: Moving past transformers for efficient open language models. *arXiv preprint arXiv:2404.07839*.

Chia-Hui Chang, Yuan-Hao Lin, and Hsiu-Min Chuang. 2020. Eventgo! exploring event dynamics from social-media posts. In *2020 International Computer Symposium (ICS)*, pages 548–552. IEEE.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Simon Gottschalk and Elena Demidova. 2018. Eventkg: A multilingual event-centric temporal knowledge graph. In *European semantic web conference*, pages 272–287. Springer.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Raphael Gruber, Abdelrahman Abdallah, Michael Färber, and Adam Jatowt. 2025. Complextempqa: A 100m dataset for complex temporal question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9111–9123.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial intelligence*, 194:28–61.

Duygu Sezen Islakoglu and Jan-Christoph Kalo. 2025. Chronosense: Exploring temporal understanding in large language models with time intervals of events. *arXiv preprint arXiv:2501.03040*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA annual convention*, volume 2, pages 1–49. Citeseer.

Zhuoyuan Liu and Yilin Luo. 2024. Document-level event extraction with definition-driven icl. *arXiv preprint arXiv:2408.05566*.

OpenAI, Aaron Hurst, Adam Lerer, and et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Gemma Team: Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, and et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Jiayu Song, Mahmud Akhter, Dana Atzil Slonim, and Maria Liakata. 2024. Temporal reasoning for timeline summarisation in social media. *arXiv preprint arXiv:2501.00152*.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

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

Qifan Wang, Bhargav Kanagal, Vijay Garg, and D Sivakumar. 2019. Constructing a comprehensive events database from the web. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 229–238.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, and et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369.

Overview of Appendices

- Appendix A: Related Work.
- Appendix B: Detailed Dataset Analysis.
- Appendix C: Prompt for Event Extraction.
- Appendix D: Prompt for FactQA question generation.

- Appendix E: Prompts for DurationQA questions.
- Appendix F: Prompts for zero-shot evaluation.
- Appendix G: More examples of English Question Answer Pairs from HistoryBankQA.
- Appendix H: Models and Compute.
- Appendix I: Human Annotation Guidelines.
- Appendix J: Detailed RAG Results.

A Related Work

Structured representations of historical events are crucial for temporal reasoning and knowledge-based tasks. Prior efforts have extracted events from semi-structured or unstructured sources. GDELT (Leetaru and Schrodt, 2013) compiles large-scale political events from news for real-time monitoring, while EventKG (Gottschalk and Demidova, 2018) builds a multilingual event knowledge graph by aligning information from structured sources like Wikidata (Vrandečić and Krötzsch, 2014), DBpedia (Auer et al., 2007), and YAGO (Hoffart et al., 2013). However, these focus on contemporary geopolitical events or rely heavily on existing ontologies. Our approach differs by aggregating historically anchored events directly from Wikipedia timelines and infoboxes, yielding fine-grained, dated events across centuries and domains. Moreover, our emphasis on multilinguality and domain diversity (science, culture, politics) distinguishes our resource from existing monolingual or domain-specific datasets.

Temporal reasoning for LLM evaluation has gained attention recently. TRAM (Wang and Zhao, 2024) introduces datasets for event ordering, arithmetic, duration, and frequency, exposing gaps between models and humans, but relies on synthetic scenarios. Other work includes graph-based reasoning (TG-LLM (Xiong et al., 2024)) and multi-level QA datasets like TempReason (Tan et al., 2023), which probes reasoning complexity but lacks multilingual or real-world contexts. TG-LLM enhances reasoning via temporal graphs, while ChronoSense (Islakoglu and Kalo, 2025) highlights LLM struggles with Allen's interval relations (Allen, 1983) on abstract and Wikidata-based events.

Concurrent to this work, three other event QA datasets have also been proposed. ComplexTempQA (Gruber et al., 2025) is a general temporal QA dataset where questions are not limited to "historical events from infobox" in the strict sense. It includes a very broad set of events, entities, and times. Compared to our HistoryBank dataset, ComplexTempQA has limited history coverage, is less multilingual and missed nuanced historical reasoning tasks (e.g., causal inference, counterfactual history). Another dataset, LexTime (Barale et al., 2025), focuses on legal events while our proposed database is very generic. In contrast, TEMPO (Abdallah et al., 2026) frames temporal reasoning primarily as a retrieval problem, evaluating a model's ability to retrieve documents that satisfy complex temporal constraints across domains. Our work instead focuses on end-to-end multilingual temporal question answering grounded in historical timelines, with deeper historical coverage and direct evaluation of long-range reasoning beyond retrieval quality alone.

While these benchmarks provide insights, they center on synthetic setups, abstract events, or everyday phenomena. In contrast, our benchmark uses factual, historically grounded events with real timestamps, enabling joint evaluation of memorization and reasoning, better reflecting real-world temporal queries. We provide a comprehensive comparison with related work in Table 7.

B Detailed Dataset Analysis

Fig. 3 shows time-wise distribution of events across all languages in HistoryBank. Data across all languages seems to follow the same trend with two notable exceptions: (1) for Italian and Russian, there is a drop from 1940s to 1950s, and (2) for French, there have been a significantly higher proportion of events published in the 2020s (until mid 2025).

Further, we also how the distribution of English

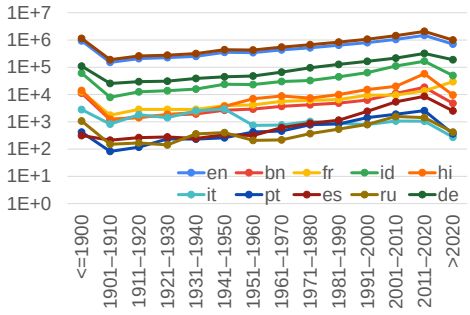

Figure 3: Time-wise Distribution of Events across all languages in HistoryBank.

Dimension	HistoryBankQA (Ours)	ComplexTempQA (Gruber et al., 2025)	TempReason (Tan et al., 2023)	ChronoSense (Islakoglu and Kalo, 2025)	TRAM (Wang and Zhao, 2024)
Domain / Focus	Multilingual temporal QA over historical events from Wikipedia timelines and infoboxes; real-world historical grounding across 10 languages.	Large-scale open-domain temporal QA requiring cross-time comparison, temporal aggregation, and multi-hop reasoning over Wikipedia/Wikidata.	Probing dataset evaluating three levels of temporal reasoning: time–time, time–event, event–event.	Benchmark evaluating LLMs on Allen's interval relations and temporal arithmetic.	Broad temporal reasoning benchmark covering ordering, arithmetic, duration, frequency, causality, and narratives across 10 datasets.
Temporal Coverage	Very deep historical coverage: 10M+ events spanning centuries to millennia.	Covers 1987–2023 with detailed metadata; primarily contemporary events.	Long-range but template-based temporal questions derived from Wikidata.	Year-level intervals from Wikidata; spans decades to centuries.	Mixed contemporary and commonsense temporal scenarios; not historically grounded.
Multilingual Support	10 languages.	English only.	English only.	English only.	English only.
Strengths / Contributions	Largest multilingual historical temporal dataset with six diverse QA tasks; strong long-range temporal reasoning; joint evaluation of factual recall and reasoning.	100M+ QA pairs (largest dataset); rich taxonomy; strong multi-hop and cross-time reasoning; detailed temporal metadata.	Systematic probing of temporal reasoning; identifies LLM biases; introduces temporal span extraction + time-sensitive RL.	First benchmark covering all 13 Allen interval relations; includes interval comparison + arithmetic; reveals weaknesses in symmetry handling.	Unified general temporal reasoning benchmark with 10 tasks; covers foundational and advanced reasoning; multiple-choice design isolates reasoning gaps.

Table 7: Comparison of our proposed HistoryBank dataset with other related datasets

events across various entity (i.e., infobox) types in Fig. 4. The distribution of English events across infobox types highlights a strong focus on biographical and cultural entities. "Officeholder" leads by a wide margin with over 1 million entries, followed by "person", "album", and "sportsperson", reflecting the prominence of political, individual, and entertainment-related content. Cultural artifacts like films, songs, and books also feature heavily, indicating rich documentation in media domains. Scientific and academic categories such as scientist, university, and academic are present but less dominant. Overall, the data suggests a strong bias toward public figures and popular culture in English-language event documentation.

Table 9 presents a list of most frequent 50 countries related to the events across languages. English (en) shows a strong Western and Anglophone focus, with the United States, United Kingdom, and Canada leading. In contrast, languages like Bengali (bn) and Hindi (hi) emphasize South Asian countries such as India, Bangladesh, and Pakistan, reflecting regional relevance. European languages like French (fr), Italian (it), and German (de) include a broader mix of European and global countries, while Indonesian (id) and Portuguese (pt) editions show strong representation from Southeast Asia and Latin America, respectively. Overall, the distribution reflects both global commonalities and regional editorial priorities in multilingual event coverage.

Fig. 5 shows polarity distribution of events across all languages in HistoryBank dataset. It reveals interesting editorial tendencies. Most languages show a strong preference for positive polarity, with Portuguese (pt) leading at over 78%, followed by French (fr) and Indonesian (id). Neutral polarity is notably higher in Hindi (hi) and Spanish (es), suggesting a more balanced or factual reporting style. Negative polarity remains relatively low across the board, with Spanish (es) being an outlier at under 1%, while Italian (it) and Russian (ru) show slightly higher negative sentiment. Overall, the dataset reflects a global inclination toward documenting events with a positive or neutral tone, with limited emphasis on negativity.

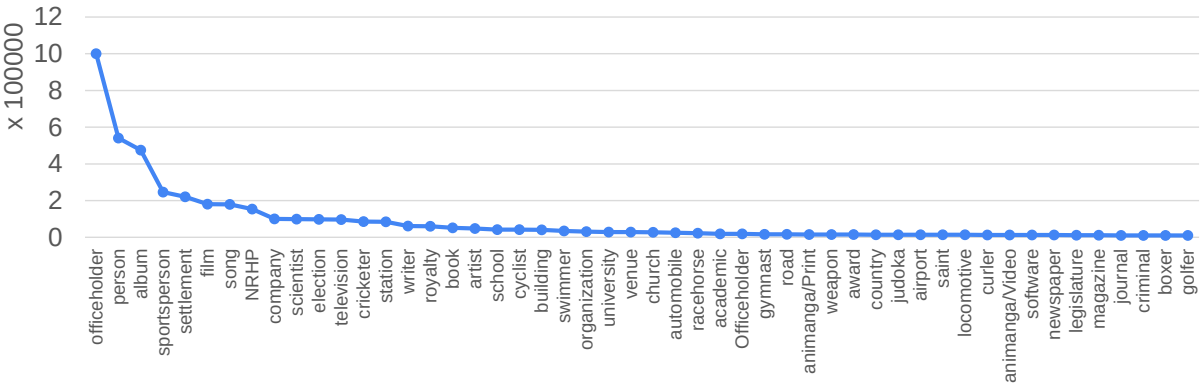

Figure 4: Entity (i.e. Infobox) type Distribution of English Events in HistoryBank. We show most frequent 50 infobox types for events from other languages in Table 8.

Lang	Most frequent 50 Infobox Types
bn	scientist, officeholder, Officeholder, royalty, person, MP, university, company, sportsperson, monarch, organization, station, election, country, book, building, television, legislature, album, Politician, cricketer, writer, spaceflight, holiday, President, software, venue, school, road, airport, philosopher, artist, award, saint, website, song, University, animanga/Video, stadium, economist, School, language, OS, bridge, magazine, president, NRHP, athlete, Country, swimmer
fr	Footballeur, Biographie2, Navire, Monument, Cycliste, Rugbyman, Politicien, Livre, Société, Sportif, Boxeur, Château, Gare, Automobile, Cheval, Saint, Athlète, Biographie, Artiste, Écrivain, Gouvernement, Cérémonie, Organisation2, Handballeur, Pièce, Assemblée, Golfeur, Art, Attentat, Stade, Pont, Route/Base, Événement, Pilote, Criminel, Arme, Musée, biographie2, Scientifique, Récompense, Cimetière, Université, Nageur, Eurovision, Blindé, Nouvelle, Logiciel, Massacre, Gratte-ciel, Exposition
id	officeholder, Officeholder, film, album, settlement, television, royalty, song, company, station, Planet, Album, spesies, scientist, sportsperson, diocese, writer, election, Film, automobile, church, President, planet, animanga/Video, animanga/Print, single, actor, Judge, artist, swimmer, Election, country, book, Politician, saint, building, award, Television, monarch, Person, venue, Airline, language, university, stadium, airport, software, weapon, Sekolah, Company
hi	settlement, cricketer, Film, television, royalty, person, book, writer, Officeholder, scientist, film, sportsperson, spaceflight, album, station, election, university, President, building, country, language, philosopher, company, OS, airport, holiday, University, legislature, organization, Person, software, stadium, artist, legislation, Judge, actor, award, venue, website, spaceflight/IP, Weapon, Company, islands, Writer, Politician, dam, airline, school, Settlement
it	nave, particella, nave/Sandbox, conflitto, cimitero, utente/Wikidipendenza, fiume, colore, album, utente, person, isola, podcast, nave/LinkCategoria, nave/Insegne, nave/Bandiera, nave/Background, malattia, lettere/cirillico, isola/Categorie, fazione militare, elemento chimico/mini tavola periodica, elemento chimico/Colore, Draft NFL, Draft NBA, Affaire criminelle
pt	person, animangá/Anime, empresa, Monarca, VG, ator, animangá/OVA, software, company, Clima, Jornal, website, haplogrupo, animangá/Mangá, television, Album, Rally, language, country, OS, carruagem, weather, atentado, University, automóvel, IFs, Biografia2, film, Weather, Sudão-Estados, mineral, casino, álbum, Société, IYPT, Company, spaceflight, settlement, eleição, Computer, Book, illustrator, Treaty, Sintetizador, Schiff/Basis, Peça, Localidade, Artefato, shopping, organization
es	company, Eurovision, NRHP, officeholder, animanga/Video, film, animanga/Print, automobile, Company, Eurovisión, television, software, person, scientist, NYCS, single, organization, VG, artist, animanga/Header, book, writer, building, filesystem, website, stadium, holiday, Organization, journal, event, animanga/Game, musical, University, Airline, boxer, flag, Software, Election, animanga/Other, OS, zoo, computer, Airport, opera, Book, wildfire, web, airport, Theatre, color
ru	animanga/Manga, software, animanga/Anime, Software, weapon, company, animanga/OVA, person, Company, planet, book, OS, referendum, animanga/Game, animanga/Film, animanga/Other, animanga/Novel, Website, wildfire, Skyscraper, animanga/Header, animanga/Movie, Weather, filesystem, animanga/Drama, Person, Indy500, biodatabase, NRHP, software, writer, Chinese, gene, Website, Software, OS, Disease, Company, winter storm small, terrorist attack, station, sports league, ship image, ship begin, rune, road/hide/tourist, road/hide/states, road/hide/ruralmuni
de	Chartplatzierungen, Leichtathlet, Unternehmen, Film, Band, Schiff, Musikalbum, Schiff/Basis, Tennisspieler, Tennisturnierjahrgang, Eishockeyspieler, Fernsehsendung, Handballer, Radsportler, Fußballsaison, Stadion, Medaillen, Schienenfahrzeug, Biathlet, Publikation, Song, Asteroid, PKW-Modell, Fußballklub, Skilangläufer, Galaxie, Basketballspieler, Verwaltungseinheit, Fluss, Flugzeug, Skispringer, Burg, Volleyballspieler, Schutzgebiet, Brücke, Gesetz, Schule, Chemikalie, Hochschule, Schwimmer, Triathlet, Schiffsklasse/Basis, Fußball-Pokalsaison, Organisation, Boxer, Bahnhof, Episode, Ort, Bauwerk, Eiskunstläufer

Table 8: Most frequent 50 infobox types for events from other languages in HistoryBank.

C Prompt for Event Extraction

```
Wikipedia has infoboxes. We have
    extracted the infobox data in a JSON
     format. Each infobox contains the
    infobox type along with the metadata
    . The metadata consists of key-value
     pairs describing the infobox. Some
    of the key-value pairs may contain
    dates and those dates might be
    related to an event. Given the
    infobox data, filter out the key-
    value pairs which have dates. Parse
    the value to extract the date in a
    year, month, day format.
    Additionally, generate an event
    description which happened on that
    date. Additionally, identify all
    possible countries to which the
    event is associated. Also output a
    polarity (positive, negative,
    neutral) for each of the possible
    countries indicating whether the
    people from the country would like
    the event. Consider if the event had
     any negative impact. Keep in mind
    that there might be multiple events
    inside a single infobox. Return the
    output as a list of JSON where each
    JSON contains the extracted date and
     event description. Each JSON should
     also contain the key-value where
```

Lang	Country List
en	United States, United Kingdom, Canada, Australia, France, India, Japan, Germany, England, Italy, Spain, Russia, China, Brazil, Ireland, Netherlands, Sweden, New Zealand, South Korea, Poland, Mexico, Iran, Scotland, Philippines, Norway, South Africa, Turkey, Argentina, Soviet Union, Ukraine, Pakistan, Belgium, Switzerland, Denmark, Hungary, Czech Republic, Greece, Malaysia, Finland, Indonesia, Austria, Romania, Portugal, Israel, Serbia, Thailand, Nigeria, Global, Chile, Croatia
bn	India, Bangladesh, United States, United Kingdom, Pakistan, England, France, Japan, Germany, China, Australia, British India, Spain, Iran, Nepal, Afghanistan, Sri Lanka, Russia, Global, Israel, Saudi Arabia, Italy, Canada, Turkey, Iraq, Soviet Union, Brazil, Egypt, Netherlands, New Zealand, Indonesia, South Africa, Argentina, Sweden, United Arab Emirates, Ottoman Empire, Syria, South Korea, Singapore, Malaysia, Thailand, Yemen, Poland, Portugal, Belgium, Ireland, Scotland, Switzerland, Ukraine, West Indies
fr	France, United States, United Kingdom, Canada, Italy, Germany, Russia, Spain, Japan, Belgium, Switzerland, South Korea, Ukraine, Netherlands, Brazil, England, Algeria, Australia, China, Soviet Union, Mexico, Morocco, Norway, Portugal, Argentina, Poland, Sweden, Ireland, Austria, Colombia, New Zealand, Denmark, Israel, Turkey, Egypt, Finland, South Africa, Chile, India, Hungary, Peru, Europe, Senegal, Croatia, Cameroon, Greece, Romania, Nigeria, Ivory Coast, Thailand, Democratic Republic of the Congo
id	Indonesia, United States, Japan, South Korea, India, France, Italy, Germany, England, United Kingdom, Russia, Global, China, Malaysia, Australia, Spain, Netherlands, Canada, Soviet Union, Brazil, Philippines, Thailand, Israel, Sweden, Poland, Taiwan, Austria, Switzerland, Hong Kong, Turkey, Denmark, Iran, Ireland, Egypt, Argentina, Singapore, Portugal, Mexico, Hungary, Belgium, Greece, Vietnam, Ukraine, Europe, Dutch East Indies, Saudi Arabia, Serbia, Norway, Romania, United Nations, South Africa, Finland, Vatican
hi	India, Pakistan, United States, Australia, England, Bangladesh, Sri Lanka, South Africa, Nepal, United Kingdom, New Zealand, Ireland, West Indies, Japan, Spain, China, Scotland, France, Canada, Zimbabwe, Germany, United Arab Emirates, Netherlands, Afghanistan, Global, Russia, British India, Thailand, Italy, Oman, Soviet Union, Uganda, Kenya, Iran, Sweden, Papua New Guinea, Britain, Namibia, Brazil, Iraq, Denmark
it	Italy, United Kingdom, United States, Japan, Germany, France, Soviet Union, Russia, Spain, Netherlands, Denmark, Finland, Canada, Austria-Hungary, Norway, Republic of Venice, Australia, Greece, Poland, Argentina, Turkey, China, India, Sweden, Austria, Croatia, Panama, South Korea, Kingdom of the Two Sicilies, Portugal, Bahamas, Ukraine, Philippines, Chile, New Zealand, West Germany, Peru, Dutch Republic, Egypt, Romania, Ottoman Empire, Brazil, Yugoslavia, England, Malta, Thailand, Not specified, Ireland, Taiwan
pt	United States, Brazil, Japan, Portugal, Germany, France, Italy, United Kingdom, South Korea, Russia, China, Australia, Canada, Spain, Israel, Sweden, Singapore, Argentina, Denmark, New Zealand, Venezuela, Soviet Union, Global, Belgium, Mexico, Norway, Austria, Uruguay, Chile, Ireland, Finland, Switzerland, Colombia, Greece, Yugoslavia, Netherlands, Turkey, England, South Africa, Serbia, Peru, Poland, Europe, Bulgaria, Cuba, Scotland, Paraguay, Azerbaijan, Ukraine, Hungary, Jamaica, Malta, Croatia
es	Japan, United States, Spain, Mexico, Brazil, Australia, United Kingdom, France, Germany, Italy, Canada, Argentina, Chile, UNASUR, South Korea, Philippines, Venezuela, Portugal, Peru, Colombia, Thailand, China, Malaysia, Taiwan, Singapore, Global, India, Saudi Arabia, Global except Asia, Indonesia, Southeast Asia, Poland, Ecuador, Panama, Hong Kong, Latin America, Netherlands, South Asia, North America, Cambodia, Austria, Laos, Europe, South Africa, Costa Rica, Dominican Republic, New Zealand, Norway, Russia, Vietnam, El Salvador
ru	United States, Japan, Russia, Germany, Soviet Union, United Kingdom, France, Italy, Canada, Global, Finland, China, Switzerland, Spain, Poland, Ukraine, Empire of Japan, Austria, Sweden, Australia, Belgium, Confederate States of America, Norway, Hungary, South Korea, Philippines, German Empire, Croatia, Czechoslovakia, Netherlands, Taiwan, Mexico, Brazil, Russian Empire, Europe, Austria-Hungary, Yugoslavia, Belarus, Argentina, Denmark, Bulgaria
de	Germany, United States, France, United Kingdom, Austria, Japan, Italy, Switzerland, Russia, Canada, Poland, Spain, Netherlands, Sweden, Australia, China, Norway, Turkey, Brazil, Czech Republic, England, Finland, Denmark, Belgium, Soviet Union, Hungary, South Korea, Romania, India, Ukraine, Peru, Argentina, Mexico, Portugal, Greece, Thailand, Slovakia, New Zealand, South Africa, Global, Croatia, Slovenia, Ireland, Colombia, Scotland, Czechoslovakia, Chile, Uruguay, Serbia, Kazakhstan, Estonia, Belarus

Table 9: List of most frequent 50 countries related to the events across languages in HistoryBank.

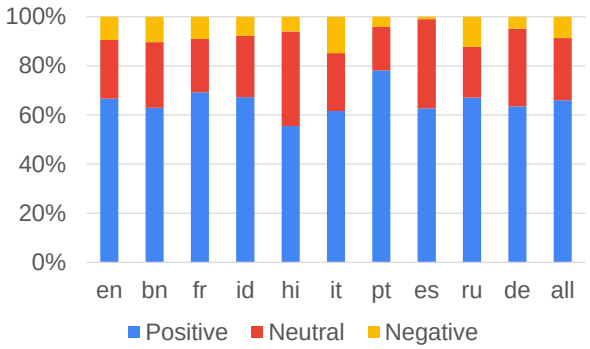

Figure 5: Polarity Distribution of Events across all languages in HistoryBank

Lang	birth	death	generic events
bn	6968	5117	6374
en	80252	37248	18941
fr	59387	35644	11979
de	93244	59863	17258
hi	1655	848	2892
id	6081	0	6276
it	0	0	14917
pt	60108	17605	14804
ru	25028	15078	18421
es	68852	28295	24995

Table 10: Data statistics per language for events mined from On This Day wikipedia pages

```
    the date came from. Do not return
    anything else other than the list.
```

Listing 1: Prompt for Event Extraction

Table 10 shows data statistics per language for events mined from On This Day wikipedia pages.

D Prompt for FactQA question generation

```
You are a helpful assistant that
    generates natural language questions
    asking for the **date of a specific
    event**.

Context: The event is described as:
- Title: <article title>
- Event Description: <event
    description>
- Infobox Type: <infobox type>
- Key: <date key>
- Language: <Language>
Your task:
- Write a fluent, natural-sounding
    **question** in <Language> that
    asks for the **date** of the
    above event.
- Use the event description and
    title to precisely identify the
    event. Avoid vague or generic
```

```
        phrasing.
    - The question must be self-
        contained, clearly referring to
        the entity or topic involved.
    - The output must be a **question
        only**. Do not include answers,
        metadata, or explanations.
    - If a valid question cannot be
        formed (e.g., due to incomplete
        or irrelevant data), return **
        INVALID**.
    Question:
```

Listing 2: Prompt for FactQA question generation

E Prompts for DurationQA questions

E.1 Prompt for checking DurationQA question validity

```
You are a knowledge validation assistant
    .

For infobox type <infobox type>,
    determine whether the following two
    events can be used to define a
    coherent duration.

- Event 1: `{k1}' - "{d1}"

- Event 2: `{k2}' - "{d2}"

Should these two events be used to form
    a valid start-end duration?

Reply with 'Yes' or 'No' only.
```

Listing 3: Prompt for checking DurationQA question validity

E.2 Prompt for DurationQA question generation

```
You are a helpful assistant that writes
    specific duration-based questions
    between two well-described events.

Context: This is about <article title>
    from a <infobox type> infobox.

Here are two events with exact
    descriptions:

1. <article title> - <start event
    description>

2. <article title> - <end event
    description>

Write a natural language specific
    question about entity (not an answer
     or a general question) that asks
    about the time duration between
    these two events.

Important:

- Refer to each event using its
    description and title exactly as
    given above and disambiguate the
    general information with exact event
     for the entity.

- Do NOT use generic questions without
    entity like -> 'How many years
    passed between the release of the
    album and the release of the single?
    '

- The question must not be general
    without entity, it is a well formed
    independent question which is clear
    without needing any external
    information.

- The output should be a **question only
    ** OR INVALID if no valid question
    or incomplete question can be formed
    .

Question:
```

Listing 4: Prompt for DurationQA question generation

F Prompts for zero-shot evaluation

```
You are an expert in history with
    extensive knowledge about past
    events. Find out the dates when each
     of the given events occurred and
    answer the question based on your
    world knowledge and reasoning.
    STRICTLY return the answer ONLY in a
     Json in the format in English:
<output>
{
"answerOut": {`year': year, `month':
    month, `day': day},
"reasoning": reasoning
}
</output>
#question#
```

Listing 5: Prompt for FactQA zero-shot inference

```
You are an expert in history with
    extensive knowledge about past
    events. Find out the dates when each
     of the given events occurred and
    answer the question based on your
    world knowledge and reasoning.
    STRICTLY return the answer ONLY in a
     Json in the format in English:
<output>
{
"answerOutStart": {`year': year, `month'
    : month, `day': day},
"answerOutEnd": {`year': year, `month':
    month, `day': day},
"reasoning": reasoning
}
</output>

#question#
```

Listing 6: Prompt for DurationQA zero-shot inference

```
You are an expert in history with
    extensive knowledge about past
    events. Find out the which event
    span is of longer duration given
    events occurred and answer the
    question based on your world
    knowledge and reasoning. STRICTLY
    return the answer ONLY in a Json in
    the format in English:
<output>
{
"answerOut": event span,
"reasoning": reasoning
}
</output>

#question#
```

Listing 7: Prompt for RelationQA Compare Duration zero-shot inference

```
You are an expert in history with
    extensive knowledge about past
    events. Find out the dates when each
     of the given event span started and
     ended and answer the question based
     on your world knowledge and
    reasoning. STRICTLY return the
    answer ONLY in a Json in the format
    in English:
<output>
{
"answerOutEventSpan1": start year-start
    month-start day/end year-end month-
    end day,
"answerOutEventSpan2": start year-start
    month-start day/end year-end month-
    end day,
"reasoning": reasoning
}
</output>

#question#
```

Listing 8: Prompt for RelationQA Duration Diff, Gap Between, and Overlap zero-shot inference

```
You are an expert in history with
    extensive knowledge about past
    events. Find out the order for which
     event group has the start date
    first than the other based on the
    starting event in each group and
    answer the question based on your
    world knowledge and reasoning. We
    are not talking about duration we
    are talking about comparison between
     starting event in each event group.
     STRICTLY return the answer ONLY in
    a Json in the format in English:
<output>
{
"answerOut": event group started first,
"reasoning": reasoning
}
</output>

#question#
```

Listing 9: Prompt for RelationQA Order Span Start zero-shot inference

```
You are an expert in history with
    extensive knowledge about past
    events. Find out the order for which
     event group has the end date later
    than the other based on the ending
    event in each group and answer the
    question based on your world
    knowledge and reasoning. We are not
    talking about duration we are
    talking about comparison between
    ending event in each event group.
    STRICTLY return the answer ONLY in a
     Json in the format in English:
<output>
{
"answerOut": event group ending last,
"reasoning": reasoning
}
</output>

#question#
```

Listing 10: Prompt for RelationQA Order Span End zero-shot inference

```
You are an expert in history with
    extensive knowledge about past
    events. Find out the which event
    group includes in itself the other
    event group based on the start and
    end date of events within each group
     and answer the question based on
    your world knowledge and reasoning.
    We are not talking about comparing
    the duration of events groups we are
     talking about inclusion of one
    event group within other group based
     on starting and ending of events
    within each group. STRICTLY return
    the answer ONLY in a Json in the
    format in English:
<output>
{
"answerOut": event group or NA,
"reasoning": reasoning
}
</output>

#question#
```

Listing 11: Prompt for RelationQA Inclusion zero-shot inference

```
You are an expert in history with
    extensive knowledge about past
    events. Find out the dates when each
     of the given events occured and
    answer the question based on your
    world knowledge and reasoning.
    Return the answer in a json in the
    format in English:
<output>
{
"answer": answer,
```

```
"reasoning": reasoning
}
</output>

#question#
```

Listing 12: Prompt for CountQA, SequenceQA and RecurrenceQA zero-shot inference

G More examples of English Question Answer Pairs from HistoryBankQA

Tables 11, 12 and 13 show multiple examples of English questions from our dataset for each question type.

H Models and Compute

LLaMA-3-8B has 8.03 billion parameters. Mistral-7B has 7.3 billion parameters. Gemma-2-9B has 9 billion parameters. Qwen3-8B has 8.2 billion parameters. Number of parameters in GPT-4o are not publicly known.

We used models only for inference. The models are accessed either using (1) API or 2) they are open source models.

All local inference experiments were run on a NVIDIA V100 node with 8 GPUs. Inference with all the models took a few hours on our test set.

I Human Annotation Guidelines

Task Overview: You will evaluate the quality of generated historical event descriptions based on information from a Wikipedia article and its infobox.

Each annotation item contains: Wikipedia article title, Infobox key–value pair(s) (source information), Generated event description

Your job is to judge how well the generated event represents the infobox information, not whether the Wikipedia article itself is correct.

Important Instructions: (1) You can use the title to search for the wikipedia page online. (2) Do not penalize minor grammatical or stylistic issues unless they affect meaning. (3) When unsure, choose the lower score.

Evaluation Criteria are as follows.

I.1 Faithfulness

Question: Does the generated event description accurately reflect the infobox information?

Scoring (3-point scale):

2 – Faithful: All facts in the description are supported by the infobox. No incorrect dates, entities, or added details. Minor rephrasing is acceptable. Example: Infobox: born = January 9, 1913 Generated event: “Richard Nixon was born on January 9, 1913.”

1 – Minor Issue: Mostly correct but with small omissions or vague wording. No clear hallucinated facts.

0 – Unfaithful: Contains hallucinated information. Incorrect date, event type, or entity. Contradicts the infobox. Example: Generated event: “Richard Nixon was elected president in 1913.”

I.2 Event Interpretability

Question: Is the event understandable and meaningful as a standalone historical event?

Scoring (3-point scale):

2 – Clear: Self-contained and specific. Understandable without additional context. Example: “The band released its debut album in 1998.”

1 – Partially Clear: Understandable but vague or awkward. Some context is missing. Example: “An album was released in 1998.”

0 – Unclear: Too generic, confusing, or incomplete. Not interpretable as an event. Example: “Something happened in 1998.”

I.3 Event Interestingness

Question: Does this infobox entry correspond to a meaningful historical “event”?

Scoring (Binary):

Yes: Birth, death, founding, election, appointment, release, merger, etc.

No: Static attributes or trivial metadata

J Detailed RAG Results

Table 14 presents the RAG results with GPT-4o as the generation model. Table 15 reports corresponding results using Mistral-7B-Instruct, LLaMA-3-8B-Instruct, Gemma-2-9B-IT, and Qwen3-8B.

Question Type	Question	Answer
FactQA	I'm looking for the date of the release of the Goodnight Punpun Omnibus 1. when was it released?	{'year': '2016', 'month': '03', 'day': '04'}
FactQA	Impax Asset Management Group plc reported £34.4 billion in assets under management in what year?	{'year': '2021', 'month': '06', 'day': '30'}
DurationQA	What was the duration between the birth and death of Scarface John Williams?	start_date {'year': '1938', 'month': '10', 'day': '19'} end_date {'year': '1972', 'month': '03', 'day': '04'}
DurationQA	What was the time duration between the release of the single "Happy Happy" and the release of the single "Fake &True"?	start_date {'year': '2019', 'month': '06', 'day': '12'} end_date {'year': '2019', 'month': '10', 'day': '18'}
CountQA - century	The following is a list of historical events. Each line includes a description and the title of the article it comes from. (1) Event about 'Sean Kingston (album)': The single "Beautiful Girls" by Sean Kingston was released. (2) Event about 'Week End (X Japan song)': The song "Week End" by X Japan was released. (3) Event about 'Lucy McEvoy': Lucy McEvoy was nominated for the AFL Women's Rising Star award. (4) Event about 'BL 5.4-inch howitzer': The Ordnance BL 5.4-inch howitzer was used in the Second Boer War. (5) Event about 'Holten Castenschiold': Holten Castenschiold began his term as the 6th President of the Danish Olympic Committee. (6) Event about 'William L. Baird': William Lewis Baird began his term as the 19th Mayor of Lynn, Massachusetts. (7) Event about 'Jung Yeon-kyung': Jung Yeon-kyung was born. Count the number of events that occurred during the 19th century. Provide the count.	2
CountQA - century	The following is a list of historical events. Each line includes a description and the title of the article it comes from. (1) Event about 'Departmental Council of Côte-d'Or': François Sauvadet was elected as President of the Departmental Council of Côte-d'Or. (2) Event about 'Constellation Place': Construction of Constellation Place began. (3) Event about 'Monroe Gooden': Monroe Washington Gooden was born. (4) Event about 'Bradenton Marauders': The Bradenton Marauders won the second half championship. (5) Event about 'Koguva': The population of Koguva was recorded as 30. List how many events fall in the 21th century. Provide the count.	3
CountQA - between_events	The following is a list of historical events. Each line includes a description and the title of the article it comes from. (1) Event about 'Moisés Paniagua': Moisés Paniagua was born. (2) Event about 'Ben Rohrer': Ben Rohrer played for Delhi Daredevils. (3) Event about 'Home Video (album)': The single 'Going Going Gone' from the album 'Home Video' was released. (4) Event about 'Black Like Me (film)': The film "Black Like Me" was released. Count the number of events that took place after event (4) but before event (2) Do not include event (4) or event (2) in the count.	1
CountQA - between_events	The following is a list of historical events. Each line includes a description and the title of the article it comes from. (1) Event about 'Buckeye Township, Michigan': Buckeye Township, Michigan was established. (2) Event about 'The Models (Mongolian TV series)': The television show first aired. (3) Event about 'Fossil Bluff': Fossil Bluff Station was established. (4) Event about 'Royal Noble Consort Uibin Seong': Ui-bin Seong began her tenure as Royal Noble Consort of the First Senior Rank. (5) Event about 'Assault Battalion No. 5 (Rohr)': Assault Battalion Number 5 (Rohr) was disbanded. What is the number of events happening strictly between event (5) and event (2)? Do not include event (5) or event (2) in the count.	2
CountQA - between_dates	The following is a list of historical events. Each line includes a description and the title of the article it comes from. (1) Event about 'Wang Dang Doodle': The song "Wang Dang Doodle" was recorded. (2) Event about '1995–96 Football League': The 1995–96 Football League Second Division season concluded. (3) Event about 'Simon Bollom': Sir Simon Bollom was born. (4) Event about 'Bahaa Taher': Bahaa Taher was awarded the Arabic Booker Prize.Count the events that occurred from 1990 to 2100. Provide the total number. Include events from both 1990 and 2100.	2
CountQA - between_dates	The following is a list of historical events. Each line includes a description and the title of the article it comes from. (1) Event about 'Franck Belot': Franck Belot was born. (2) Event about 'Sorato Anraku': Sorato Anraku won a gold medal in Bouldering at the Innsbruck 2023 IFSC Climbing World Cup. (3) Event about 'SV Babelsberg 03': SV Babelsberg 03 was founded. (4) Event about 'Barack Obama vs. Mitt Romney (video)': The previous song "Frank Sinatra vs. Freddie Mercury" was released. (5) Event about '1919 Ayvalık earthquake': The 1919 Ayvalık earthquake occurred, causing significant destruction and loss of life. (6) Event about 'Mananciais do Rio Paraíba do Sul Environmental Protection Area': The Mananciais do Rio Paraíba do Sul Environmental Protection Area was created. (7) Event about 'Akshay Venkatesh': Akshay Venkatesh was awarded the Infosys Prize. Within the time span of 1920–2050, how many of these events occurred? Provide the total number. Include events from both 1920 and 2050.	5

Table 11: Examples of Question-Answers pairs for FactQA, DurationQA and CountQA question types in English

Question Type	Question	Answer
Compare Duration	Which event span is of longer duration: “Agnes van Ardenne was born” to “Agnes van Ardenne started her earlier term as Member of the House of Representatives”, or “Agnes van Ardenne ended her earlier term as Member of the House of Representatives” to “Agnes van Ardenne started her term as Member of the House of Representatives.” in the context of “Agnes van Ardenne”?	“Agnes van Ardenne was born” to “Agnes van Ardenne started her earlier term as Member of the House of Representatives.”
Compare Duration	Which event span is of longer duration: “Albert Planasdemunt i Gubert began his term as a Member of the Parliament of Catalonia” to “Albert Planasdemunt i Gubert began his term as Mayor of Breda”’, or “Albert Planasdemunt i Gubert was born” to “Albert Planasdemunt i Gubert passed away.” in the context of “Albert Planasdemunt i Gubert”?	“Albert Planasdemunt i Gubert was born” to “Albert Planasdemunt i Gubert passed away.”
Duration Diff	What is the difference in duration between “A new session of the Alabama Legislature began” to “The next election for the Alabama Senate is scheduled”, and “Anthony Daniels was elected as House Minority Leader of Alabama” to “The last election for the Alabama Senate was held.” (in days) in the context of “Alabama Legislature”?	762
Duration Diff	What is the difference in duration between “Adolf von Thadden was born” to “Adolf von Thadden began his term as a Member of the Bundesrat.” and “Adolf von Thadden ended his term as a Member of the Bundestag” to “Adolf von Thadden ended his term as a Member of the Bundesrat.” (in days) in the context of “Adolf von Thadden”?	10280
Gap Between	What is the gap in days between “José Luis Soro became the leader of Chunta Aragonesista” to “Maru Díaz became the leader of Podemos–Green Alliance in Aragon.”, and “Tomás Guitarte became the leader of Teruel Existe” to “Alberto Izquierdo became the leader of the Aragonese Party.”, in the context of “2023 Aragonese regional election”?	1522
Gap Between	What is the gap in days between “The January 2007 special session of the 98th Wisconsin Legislature started” to “The March 2008 special session of the 98th Wisconsin Legislature started”, and “The December 2007 special session of the 98th Wisconsin Legislature ended” to “The April 2008 special session of the 98th Wisconsin Legislature ended.”, in the context of “98th Wisconsin Legislature”?	62
Inclusion	Which event’s time span includes the other: “June 2009 Extra Session of the 99th Wisconsin Legislature ended” to “December 2009 Special Session of the 99th Wisconsin Legislature began”, or “Election for the 99th Wisconsin Legislature was held” to “The term of the 99th Wisconsin Legislature ended.”, in the context of “99th Wisconsin Legislature”?	“Election for the 99th Wisconsin Legislature was held” to “The term of the 99th Wisconsin Legislature ended.”
Inclusion	Which event’s time span includes the other: “Ahmad Maslan ended his term as Deputy Minister of Finance” to “Ahmad Maslan began his term as State Deputy Chairman of the United Malays National Organisation of Johor”, or “Ahmad Maslan ended his term as Deputy Minister of International Trade and Industry” to “Ahmad Maslan began his term as Secretary-General of Barisan Nasional.”, in the context of “Ahmad Maslan”?	“Ahmad Maslan ended his term as Deputy Minister of Finance” to “Ahmad Maslan began his term as State Deputy Chairman of the United Malays National Organisation of Johor.”
Order Span End	Which event span ended last: “The album ‘100 Reasons to Live’ by Gareth Emery was released” to “The single ‘Far From Home’ by Gareth Emery feat. Gavrielle was released”, or “The single ‘Hands’ by Gareth Emery &Alastor feat. London Thor was released” to “The single ‘Save Me’ by Gareth Emery was released.”, in the context of “100 Reasons to Live”?	“The single ‘Hands’ by Gareth Emery &Alastor feat. London Thor was released” to “The single ‘Save Me’ by Gareth Emery was released”.
Order Span End	Which event span ended last: “The March 2018 extraordinary session of the 103rd Wisconsin Legislature started” to “The 103rd Wisconsin Legislature term ended”, or “The January 2018 special session of the 103rd Wisconsin Legislature started” to “The March 2018 extraordinary session of the 103rd Wisconsin Legislature ended.”, in the context of “103rd Wisconsin Legislature”?	“The March 2018 extraordinary session of the 103rd Wisconsin Legislature started” to “The 103rd Wisconsin Legislature term ended.”
Order Span Start	Which event span started earlier: “The album ‘100 Reasons to Live’ by Gareth Emery was released” to “The single ‘Far From Home’ by Gareth Emery feat. Gavrielle was released”, or “The single ‘Hands’ by Gareth Emery &Alastor feat. London Thor was released” to “The single ‘Save Me’ by Gareth Emery was released.”, in the context of “100 Reasons to Live”?	“The single ‘Hands’ by Gareth Emery &Alastor feat. London Thor was released” to “The single ‘Save Me’ by Gareth Emery was released.”
Order Span Start	Which event span started earlier: “The March 2018 extraordinary session of the 103rd Wisconsin Legislature started” to “The 103rd Wisconsin Legislature term ended”, or “The January 2018 special session of the 103rd Wisconsin Legislature started” to “The March 2018 extraordinary session of the 103rd Wisconsin Legislature ended.”, in the context of “103rd Wisconsin Legislature”?	“The January 2018 special session of the 103rd Wisconsin Legislature started” to “The March 2018 extraordinary session of the 103rd Wisconsin Legislature ended.”
Overlap	How many days do “The single ‘Hands’ by Gareth Emery &Alastor feat. London Thor was released” to “The single ‘Far From Home’ by Gareth Emery feat. Gavrielle was released.”, and “The album ’100 Reasons to Live’ by Gareth Emery was released” to “The single ’Save Me’ by Gareth Emery was released.” overlap, in the context of “100 Reasons to Live”?	29
Overlap	How many days do “The 2016 Wisconsin elections were held” to “The November 2018 extraordinary session of the 103rd Wisconsin Legislature started.”, and “The January 2017 special session of the 103rd Wisconsin Legislature ended” to “The March 2018 special session of the 103rd Wisconsin Legislature started.” overlap, in the context of “103rd Wisconsin Legislature”?	275

Table 12: Examples of Question-Answers pairs for RelationQA question type in English

Question type	Question	Answer
SequenceQA - Verify	Is the following timeline historically valid? Each event includes a description and the title of the article it came from. Answer 'True' if the events are in correct chronological order, otherwise 'False'. (1) Event about 'Mudhoji II of Nagpur': Mudhoji II began his reign as the 6th Raja of Nagpur. (2) Event about 'L'Épiphanie': L'Épiphanie was officially constituted as a city. (3) Event about 'Dossena': The population of Dossena was recorded as 966. (4) Event about 'Next (Indian retailer)': The company profile page of Next Retail India Ltd was archived. (5) Event about 'Richard Money': Richard Money began managing Hartlepool United.	FALSE
SequenceQA - Verify	Do the events occur in proper chronological sequence? Each event includes a description and the title of the article it came from.Answer 'True' if the events are in correct chronological order, otherwise 'False'. (1) Event about 'River Vale Skeeters': The River Vale Skeeters hockey team began operations. (2) Event about 'Sandrine Hamel': Sandrine Hamel won a bronze medal in Dual banked slalom at the 2021 World Para Snow Sports Championships in Lillehammer. (3) Event about 'Dominique Walter': Dominique Walter was born. (4) Event about 'Castaic Dam': Castaic Dam was opened.	FALSE
SequenceQA - MCQ	Only one of the sequences below is in the correct order. Can you find it? Each event is described alongside its article title. Choose the correct order using the event descriptions and article titles provided. (A) 'Zaza Nadiradze': Zaza Nadiradze was born. | 'ND Slovan': Nogometno društvo Slovan was founded. | 'Daniele Rimpelli': Daniele Rimpelli represented the Italy national rugby union team. | 'Matt Viney': Matt Viney began his term as a Member of the Victorian Legislative Council for Eastern Victoria Region. (B) 'ND Slovan': Nogometno društvo Slovan was founded. | 'Daniele Rimpelli': Daniele Rimpelli represented the Italy national rugby union team. | 'Matt Viney': Matt Viney began his term as a Member of the Victorian Legislative Council for Eastern Victoria Region. | 'Zaza Nadiradze': Zaza Nadiradze was born. (C) 'ND Slovan': Nogometno društvo Slovan was founded. | 'Zaza Nadiradze': Zaza Nadiradze was born. | 'Matt Viney': Matt Viney began his term as a Member of the Victorian Legislative Council for Eastern Victoria Region. | 'Daniele Rimpelli': Daniele Rimpelli represented the Italy national rugby union team. (D) 'Matt Viney': Matt Viney began his term as a Member of the Victorian Legislative Council for Eastern Victoria Region. | 'Zaza Nadiradze': Zaza Nadiradze was born. | 'Daniele Rimpelli': Daniele Rimpelli represented the Italy national rugby union team. | 'ND Slovan': Nogometno društvo Slovan was founded.	C
SequenceQA - MCQ	Select the option that shows the correct chronological order of events. Each event is described alongside its article title.Choose the correct order using the event descriptions and article titles provided. (A) 'Darrel Higham': Darrel Higham was born. | 'Type 620 tanker': The ship Shengli had its maiden voyage. | 'Sian Williams': Sian Williams married Neale Hunt. (B) 'Type 620 tanker': The ship Shengli had its maiden voyage. | 'Darrel Higham': Darrel Higham was born. | 'Sian Williams': Sian Williams married Neale Hunt. (C) 'Type 620 tanker': The ship Shengli had its maiden voyage. | 'Sian Williams': Sian Williams married Neale Hunt. | 'Darrel Higham': Darrel Higham was born. (D) 'Darrel Higham': Darrel Higham was born. | 'Sian Williams': Sian Williams married Neale Hunt. | 'Type 620 tanker': The ship Shengli had its maiden voyage.	A
SequenceQA - Arrange	Sort the events by the time they took place. Each event is accompanied by a description and the title of the article it comes from. Return the correct chronological order by listing the event numbers, like (2) (1) (3). (1) Event about 'Paaliaq': Paaliaq was discovered. (2) Event about 'Bob Chakales': Bob Chakales was born. (3) Event about 'Dark Valley': The film "Dark Valley" was released.	(2) (3) (1)
SequenceQA - Arrange	Given the events below, provide their correct order in history. Each event is accompanied by a description and the title of the article it comes from. Return the correct chronological order by listing the event numbers, like (2) (1) (3). (1) Event about 'Barnstaple Town railway station': Barnstaple Town station was closed. (2) Event about 'RTHK TV 31': RTHK TV 31 closed its analogue service. (3) Event about 'Monade': Monade disbanded. (4) Event about 'Order of the African Star': The Order of the African Star became a Belgian Order.	(4) (1) (3) (2)
RecurrenceQA	The following event includes the article title and event description. Event about '2008 Iranian legislative election': The second round of the 2008 Iranian legislative election was held. Identify the year when the previous edition of the event took place. Provide the year as your answer.	2004
RecurrenceQA	The following event includes the article title and event description. Event about '2024 Green National Convention': The Green Party National Political Convention began. When was the previous edition of this event? Provide the year as your answer.	2020

Table 13: Examples of Question-Answers pairs for SequenceQA and RecurrenceQA question types in English

Task		en	bn	de	fr	id	hi	it	pt	ru	es
FactQA		33.7	39.8	36.1	35.5	53.5	44.1	29.0	49.2	47.8	42.5
DurationQA		22.0	65.5	73.9	59.5	79.3	68.5	70.5	65.3	74.5	59.6
RelationQA	Compare Duration	50.6	45.3	64.8	46.3	68.5	62.0	65.5	55.1	-	41.7
	Duration Diff	9.5	5.3	25.2	7.1	29.0	20.4	15.5	32.4	-	10.7
	Gap Between	9.1	5.4	21.3	5.1	31.6	20.4	-	-	-	-
	Inclusion	48.9	6.7	67.2	29.2	49.2	58.5	-	-	-	-
	Order Span End	41.5	47.3	48.9	41.2	60.1	60.1	-	-	-	27.0
	Order Span Start	58.2	64.2	74.1	54.4	71.3	71.3	-	-	-	60.9
	Overlap	7.5	4.9	24.6	7.5	17.0	17.7	15.9	31.1	-	12.0
CountQA		31.1	27.5	33.7	32.2	35.3	34.5	31.6	37.3	35.7	37.1
SequenceQA		30.9	39.9	40.7	33.3	46.9	41.9	42.3	43.6	49.1	35.4
RecurrenceQA		84.7	91.3	-	-	93.5	97.3	-	-	-	-

Table 14: RAG Results with GPT4o across languages on reasoning tasks in our proposed HistoryBankQA benchmark. We don't report results for cells with test sample size < 30.

Task		Mistral-7B-Instruct										LLaMA-3-8B-Instruct									
		en	bn	de	fr	id	hi	it	pt	ru	es	en	bn	de	fr	id	hi	it	pt	ru	es
FactQA		22.1	11.2	20.5	19.3	23.8	20.4	17.6	22.7	21.4	19.7	15.8	9.1	13.2	12.3	16.7	13.9	10.2	15.1	14.4	12.9
DurationQA		19.7	53.1	56.4	50.2	58.5	55.1	52.8	55.9	56.4	50.9	14.1	27.8	30.1	27.3	31.5	28.9	26.7	29.6	30.1	27.1
RelationQA	Compare Duration	27.2	18.4	29.3	25.2	31.6	28.1	26.3	30.4	-	24.7	19.8	13.4	21.1	17.4	22.8	20.3	18.9	22.2	-	16.8
	Duration Diff	6.8	2.5	10.8	4.7	13.2	9.4	8.1	15.1	-	5.6	4.1	1.7	5.9	2.3	7.2	5.3	4.8	7.7	-	3.2
	Gap Between	6.2	2.3	8.5	4.1	13.9	9.4	-	-	-	-	3.7	1.2	4.6	1.9	8.1	5.2	-	-	-	-
	Inclusion	32.2	4.2	38.7	23.2	40.5	36.4	-	-	-	-	20.8	2.5	24.2	14.1	27.8	21.8	-	-	-	-
	Order Span End	33.7	31.1	37.3	32.8	49.1	48.7	-	-	-	23.9	23.2	19.4	25.6	21.9	32.7	32.1	-	-	-	14.7
	Order Span Start	47.1	42.6	51.2	45.4	59.7	59.2	-	-	-	44.9	34.1	30.2	37.3	32.5	43.9	43.3	-	-	-	33.1
	Overlap	5.4	1.8	9.6	4.2	11.8	11.1	9.3	18.2	-	7.1	2.9	1.1	6.2	2.5	7.1	6.9	5.3	12.1	-	4.3
CountQA		17.7	2.3	2.5	0.3	4.7	0.5	13.7	14.4	2.5	13.3	12.6	1.7	0.9	2.8	1.3	0.8	12.9	13.5	4.7	14.7
SequenceQA		6.9	2	1.1	6.6	1.2	3.3	7.1	3.1	2.4	7	15.2	2.6	3.5	12.6	1.1	1.9	7.3	15.7	10	19.1
RecurrenceQA		44.6	5.2	-	-	56.2	8.9	-	-	-	-	40.4	2.9	-	-	53.4	35.3	-	-	-	-

Task		Gemma-2-9B-IT										Qwen3-8B									
		en	bn	de	fr	id	hi	it	pt	ru	es	en	bn	de	fr	id	hi	it	pt	ru	es
FactQA		29.5	23.6	31.1	28.9	35.6	31.4	26.1	34.8	34.1	31.2	27.1	17.3	28.4	26.3	32.8	28.7	24.1	32.2	31.6	28.4
DurationQA		31.1	64.2	67.8	61.7	70.1	66.2	63.5	67.1	67.6	62.4	27.8	59.6	62.9	57.2	64.8	61.1	58.6	62.3	62.8	57.9
RelationQA	Compare Duration	41.2	33.2	45.4	38.4	48.7	45.2	43.1	47.3	-	37.9	36.8	26.4	40.1	34.3	43.6	40.1	37.9	42.1	-	33.3
	Duration Diff	12.2	6.2	18.9	10.1	22.1	17.5	15.4	24.8	-	11.4	10.7	4.6	16.4	8.7	19.2	14.8	13.4	21.7	-	9.5
	Gap Between	10.9	5.1	15.6	8.6	25.1	19.1	-	-	-	-	9.4	3.7	13.2	7.2	21.6	16.3	-	-	-	-
	Inclusion	47.9	11.9	57.4	35.4	60.1	54.3	-	-	-	-	44.2	9.2	53.1	32.1	55.7	49.6	-	-	-	-
	Order Span End	48.2	44.6	52.1	45.6	63.6	62.8	-	-	-	31.2	42.7	37.9	46.2	40.3	56.7	56.1	-	-	-	27.7
	Order Span Start	61.8	57.3	67.8	60.2	74.9	74.3	-	-	-	56.4	56.4	51.6	61.9	55.3	68.3	67.6	-	-	-	51.5
	Overlap	9.1	3.7	16.8	8.3	19.3	18.2	15.3	25.8	-	11.9	7.9	2.7	14.2	6.8	16.1	15.2	12.8	21.4	-	9.9
CountQA		3.5	1.1	3.3	3.2	2.4	1.2	6.1	3	2.3	2.3	2.5	0.2	0	0.3	0.3	0.3	2	1.7	0.1	6.9
SequenceQA		3.9	1.3	3.6	3.8	2.9	1	4.5	4.4	3.8	1.4	2.5	0.8	0.1	2.2	0.1	1	1.8	2.9	1.4	5.9
RecurrenceQA		15.2	26.2	-	-	26.7	31	-	-	-	-	7.2	4.7	-	-	8.7	17.1	-	-	-	-

Table 15: Mistral-7B-Instruct (top left), LLaMA-3-8B-Instruct (top right), Gemma-2-9B-IT (bottom left), Qwen3-8B (bottom right) RAG performance across languages on temporal reasoning tasks in our proposed HistoryBankQA benchmark. We don't report results for cells with test sample size less than 30.