HISTOBENCH: WORLD HISTORY EVENT EXTRACTION AND COGNITIVE-LEVEL BENCHMARKING OF GENERATIVE AI

Anonymous authors
Paper under double-blind review

ABSTRACT

We present HistoBench, a benchmark and dataset designed to evaluate and improve large language models' (LLMs) ability to reason about complex, temporally grounded historical narratives. While LLMs perform well on general language tasks, their historical understanding remains limited. HistoBench provides a richly annotated collection of global events, timelines, and causal chains, alongside an interactive timeline and global map to enhance accessibility for research and education. To assess reasoning across multiple depths, we introduce a set of 1,007 historical questions structured around Bloom's Taxonomy, covering levels from factual recall (*Remember*) to higher-order reasoning (*Evaluate* and *Create*). Our results show that models perform well on spatial and entity recognition but struggle more with temporal reasoning. Among the evaluated systems, DeepSeek-V3 consistently outperforms GPT4o-mini and Gemma-3 across nearly all levels, achieving over 90% accuracy at the most advanced stages of evaluation and creation, highlighting its stronger capacity for complex historical reasoning.

1 INTRODUCTION

The emergence of digital humanities over the last two decades has fundamentally transformed scholarship in the humanities, particularly in the field of history (Fafalios et al., 2023). Historical documents are now massively digitized into photos and texts, allowing researchers to query across collections and languages. This digitization has created an enormous volume of archives and archival data available digitally, producing a valuable but under-utilized source of large-scale digital data for digital humanities scholars (Hawkins, 2021). However, several challenges remain in this domain.

The challenges in the historical data analysis are as follows: **(i) Under-exploration of certain historical tasks:** One of the primary challenges in digital humanities is the under-exploration of certain historical tasks, particularly event extraction, which has either been applied to small-scale datasets or constrained by limited event typologies with predefined event categories (Rovera et al., 2019) (Hervieux et al., 2024). This limitation has restricted the broader application and generalizability of event extraction methods in historical research. **(ii) The lack of structured data:** Most historical texts are not in clean, structured formats suitable for direct computational analysis, therefore requiring extensive preprocessing before being usable in NLP pipelines (Wakabayashi, 2019). Available historical texts can be divided into three types from the point of automated text analysis: initially digital, printed/written but digitized, and non-digitized printed/written texts (Huistra & Mellink, 2016). In the case of solely printed or written texts, digitization is just the first step, as digitized text must be preprocessed to make it proper for automated analysis through steps like correction of Optical Character Recognition (OCR), concept or meta tagging, and lemmatization (Szabó et al., 2020). **(iii) Presenting large historical datasets:** While large-scale analysis of historical sources can provide a broader and more nuanced understanding of historical events, the sheer volume of extracted data can be overwhelming. For it to be useful, especially to non-experts, the data must be organized, filtered, and displayed in an accessible and user-friendly format. The scale and diversity of such collections presents particular challenges in identifying and extracting relevant content (Leavy et al., 2019). **(iv) Benchmarking Gaps in Historical Knowledge Evaluation:** Evaluating large language models on historical knowledge has become a key area of research as these systems are increasingly used for educational and informational purposes (Garcia & Weilbach, 2023). History presents unique challenges for LLMs because it requires not just memorizing isolated facts, but understanding complex relationships between events, people, and time periods (Kandpal et al., 2023). Moreover, our historical knowledge and the available digital data are heavily skewed toward Western narratives, and this Western bias is also evident in the knowledge encoded by large language models (Keleg & Magdy, 2023).

To address the first challenge, we employed large language models (LLMs) and used prompt engineering techniques to perform tasks such as historical event extraction. To tackle the second challenge, we developed a series of preprocessing steps, particularly tailored to the constraints and nuances of feeding book-length texts into LLMs. To overcome the third challenge, we designed a web-based user interface that enables users to visually explore and filter the extracted events through interactive timelines and maps. Therefore, both academic researchers and non-specialist users can benefit from the outputs. Scholars can use the platform for historical investigations across a wide range of time periods and geographic regions, regardless of their specific area of expertise. In addition, the platform serves as an educational tool, accessible to general users with an interest in learning about historical events and patterns. To address the forth gap, we curated a dataset of 1,007 multiple-choice questions derived from the structured historical data extracted from our source texts. This dataset covers a wide variety of time periods and regions, enabling a fair and representative evaluation. We then used it to benchmark the historical understanding of several state-of-the-art LLMs, providing new insights into their performance and limitations in processing historical content. Figure 1 provides a visual overview of the event extraction process and large language model (LLM) evaluation pipelines in our work.

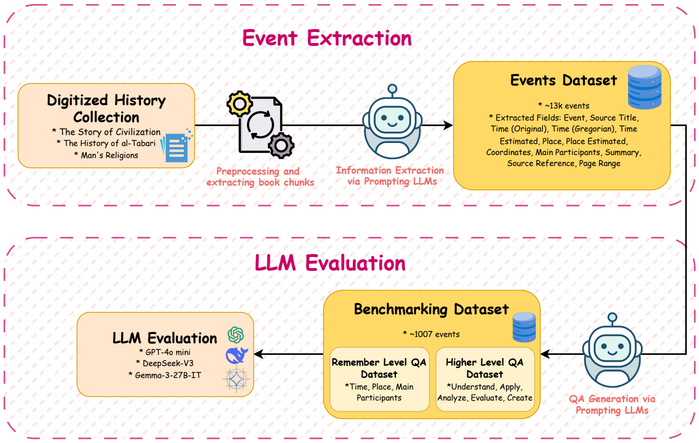

Event Extraction
Digitized History Collection
* The Story of Civilization
* The History of al-Tabari
* Man's Religions
Preprocessing and extracting book chunks
Information Extraction via Prompting LLMs
Events Dataset
* ~13k events
* Extracted Fields: Event, Source Title, Time (Original), Time (Gregorian), Time Estimated, Place, Place Estimated, Coordinates, Main Participants, Summary, Source Reference, Page Range
LLM Evaluation
QA Generation via Prompting LLMs
Benchmarking Dataset
* ~1007 events
Remember Level QA Dataset
*Time, Place, Main Participants
Higher Level QA Dataset
*Understand, Apply, Analyze, Evaluate, Create
LLM Evaluation
* GPT-4o mini
* DeepSeek-V3
* Gemma-3-27B-IT

Figure 1: An overview of our pipeline for historical event extraction and evaluation. The top section illustrates how structured event data is extracted from digitized historical texts using LLMs. The bottom section shows how the resulting dataset is used for evaluating LLMs across multiple reasoning levels based on Bloom's Taxonomy.

2 Related Work

Event Extraction: A common approach in the task of event extraction has been to decompose it into smaller subtasks. For example, (Nguyen & Grishman, 2018) employs Graph Convolutional Networks (GCNs) to perform event detection, which involves identifying whether a specific event occurs within a given text. Another example is GRIT (Du et al., 2021), which uses a transformer-based model to extract entities related to events.

Subsequent work in event extraction has largely framed the task as a classification problem, often focusing on identifying and categorizing event triggers—words that explicitly indicate the occurrence of an event, typically the main verb in a sentence. This approach is based on annotation guidelines such as those provided by the ACE dataset (ACE), which defines and categorizes event types. For example, Sprugnoli and Tonelli (Sprugnoli & Tonelli, 2019) introduced an annotation scheme that classifies events into 22 categories and created a dataset with these annotations, along with a model to automate the annotation process. The BRAD dataset (Lai et al., 2021) is another relevant example. It contains annotated historical texts related to Black uprisings found in 19th-century African American newspapers. The study reported that existing models, based primarily on BERT, struggled to perform well on this dataset.

A significant shift in methodology came with research showing that framing event extraction as a question answering (QA) task yields promising results [liu-etal-2020-event]. Follow-up studies have validated the effectiveness of this approach. For instance, (Borenstein et al., 2023) introduced a multilingual dataset based on early modern colonial-era newspaper advertisements that document formerly enslaved individuals who liberated themselves. Using a QA-based approach with RoBERTa models, they achieved strong results on these historical texts.

However, these prior works have notable limitations: the questions are typically handcrafted, the tasks are limited to specific event types, and the datasets are small in scale and narrowly focused. Given the demonstrated success of QA formulations for event extraction, the emergence of large language models (LLMs) presents a powerful opportunity. These models inherently operate well in QA-like formats and enable large-scale, high-accuracy event extraction across diverse historical texts, without being constrained by fixed event taxonomies.

Visualizing Historical Events: In terms of visualizing historical events on a timeline, relatively few studies have addressed this challenge. Bedi et al. (Bedi et al., 2017) utilized the TimeMapper tool (https://timemapper.okfnlabs.org/) for this purpose, using the NER component of Stanford CoreNLP (Manning et al., 2014) to extract events. However, their extracted events were limited in scope, based on only around 200 sentences. Another study by Hienert et al. (Hienert & Luciano, 2012) worked with a larger dataset spanning from 300 BC to 2013. Their dataset was derived from structured data on Wikipedia, where events are already listed in chronological format on dedicated pages. Their work focused primarily on building a pipeline for event extraction and visualization from this semi-structured source.

Historical Benchmarking for LLMs: General-purpose evaluation benchmarks like MMLU (Hendrycks et al., 2021) are widely adopted across numerous academic domains, including history, as proxies for assessing large language models' reasoning and encyclopedic knowledge. However, these benchmarks are not tailored to the unique demands of historical reasoning: they do not offer contextual narrative structure, causal chaining, or temporally grounded evaluation specific to history, motivating the need for a domain-specific dataset.

Dedicated historical and temporal reasoning benchmarks have made important progress, but each exhibits key limitations. HiST-LLM, built from the Seshat Global History Databank, provides structured coverage of historical societies from the Neolithic to the Industrial Revolution, but emphasizes basic factual recall and lacks systematic alignment with cognitive levels like analysis or evaluation (Hauser et al., 2024). HistBench, developed alongside the HistAgent platform, offers multilingual and multimodal historical QA, yet remains limited in scale (hundreds of questions) and does not integrate Bloom's Taxonomy to balance cognitive complexity across tasks (Qiu et al., 2025). Temporal reasoning benchmarks such as TRAM (Wang & Zhao, 2024) and TimeBench (Chu et al., 2024) provide broad coverage of tasks involving ordering, duration, frequency, arithmetic, and some aspects of causality. Nonetheless, they lack support for causal-chain visualizations and structured narrative event extraction, and similarly omit a systematic approach to cognitive-level design.

In contrast, our work addresses these gaps by delivering (1) broad temporal and geographic representation of extracted events; (2) an interactive, map-based visualization interface; and (3) a deliberately designed set of 1,007 multiple-choice questions, crafted according to Bloom's Taxonomy to span remembering through creating cognitive levels. This enables more interpretable and cognitively informed evaluation of LLM historical reasoning.

Table 1: Basic quantitative statistics of the selected historical texts, including total pages, word counts, and character counts

Book	**Pages #**	**Words #**	**Characters #**
The Story of Civilization	9,570	4.24M	24.7M
The History of al-Tabari	6,166	1.63M	8.11M
Man's Religions	768	0.30M	1.76M
Total	**16,504**	**6.17M**	**34.6M**

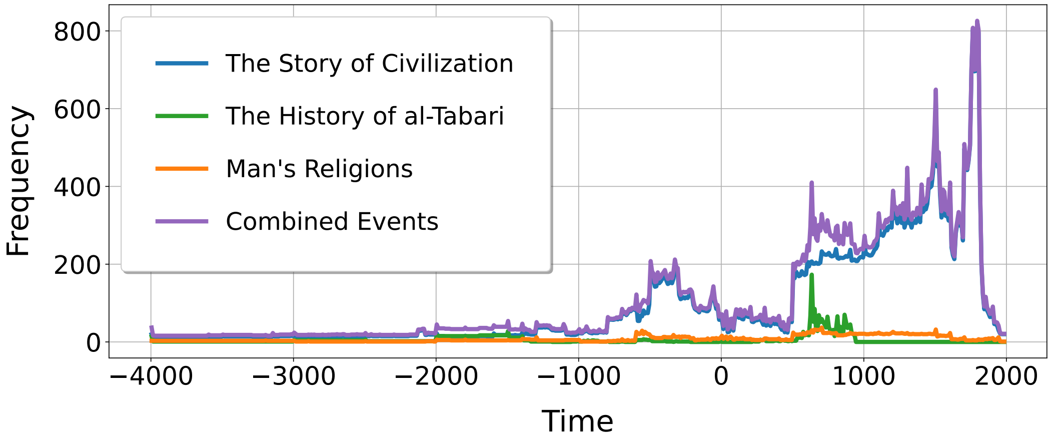


Figure 2: Temporal distribution of events in the full dataset, categorized by source texts: The Story of Civilization, The History of al-Tabari, and Man's Religions.

3 Dataset

3.1 Events Dataset

We analyzed three major historical texts [1] to extract a wide range of world events, aiming to broaden the geographic and cultural scope beyond a predominantly Western focus. Some information about the size of these resources is provided in Table 1, which summarizes the number of pages, words, and characters for each book as well as their combined totals. Our resources include:

The Story of Civilization, an 11-volume series by Will and Ariel Durant (1935–1975), traces the broad sweep of world history from prehistoric times through the Napoleonic era in 1975. While it covers both Eastern and Western civilizations, the narrative foregrounds European and Western developments, weaving together political, cultural, and intellectual histories with storytelling for a general readership (Durant, 1942). For detailed volume-specific distributions, see Figure 5 (temporal distribution of events) in the Appendix.

The History of al-Tabari (also known as *Tarikh al-Rusul wa al-Muluk*), compiled by Ibn Jarir al-Tabari and completed in 915CE, is an 11-volume annalistic chronicle beginning with creation and covering ancient empires, prophetic traditions, and Islamic history through to the early Abbasid caliphate. It offers an in-depth account of Middle Eastern history up to 915CE, with particular emphasis on Persian and early Islamic narratives (al Tabari & Rosenthal, 1988). The original text is in Arabic, and we conducted our analysis directly on the Arabic version to avoid potential issues introduced by translation nuances.

[1] We used three major historical works: *The Story of Civilization*, *The History of al-Tabari*, and *Man's Religions*, to enrich our dataset. No copyrighted text was reproduced; all historical content was paraphrased and fully attributed. This use aligns with standard academic fair-use (U.S.) and fair-dealing (U.K. and similar jurisdictions) practices, which permit paraphrasing factual material for non-commercial scholarly research provided attribution is given and no substantial portions of original expression are copied.

Man's Religions (by John B.Noss; revised edition c.1980s; originally early 1960s editions) is a single-volume comparative overview of global belief systems. It is organized in four thematic sections: primitive and extinct religions, religions of India, religions of East Asia, and religions of the Middle East, and provides factual, comparative descriptions of each tradition's history, beliefs, and practices (Noss, 1956).

Drawing on these sources and after the aggregation process, the resulting dataset includes **13,233 historical events**, categorized as follows: 11,176 from *The Story of Civilization*, 1,570 from *The History of al-Tabari*, and 487 from *Man's Religions*. The temporal distribution of these events is illustrated in Figure 2, which shows a higher density in the last 1,500 years. Each extracted event in our dataset is represented using the structured fields detailed in Table 2.

Table 2: Universal data schema for historical events

Field	**Description**
Event	A short title or description of the event
Source title	Title of the event as it appears in the original text (if applicable)
Time (original)	Temporal description of the event as provided by the source
Time (gregorian)	Normalized year in the Gregorian calendar (negative for BCE, positive for CE)
Time estimated	Boolean flag: `true` if inferred, `false` if explicitly given in the source
Place	Name of the geographical location where the event occurred
Place estimated	Boolean flag: `true` if inferred, `false` if stated in the source
Coordinates	Standardized latitude and longitude of the location
Main participants	Key individuals or groups involved in the event
Summary	A concise summary of the event, optionally generated by a language model
Source reference	Name and volume of the source
Page range	Start and end pages of the event in the source material

3.2 Benchmarking Dataset

To evaluate the performance of large language models (LLMs), we constructed a balanced benchmarking subset derived from our large-scale event dataset.

3.2.1 Event Selection

From the full corpus of 13,233 historical events, we selected a representative subset of 1,007 instances, ensuring coverage across diverse geographic regions, historical periods, and thematic domains. The dataset size was intentionally limited to a scale feasible for manual verification, thereby supporting the correctness and reliability of the benchmark. The distribution of the selected events is visualized in Figure 3, which demonstrates a similar distribution pattern between the full dataset and the benchmarking subset. Events from earlier historical periods are depicted in blue, transitioning to red for more recent events. Furthermore, areas with greater event density are represented with more intense colors, highlighting regions of significant historical concentration.

3.2.2 Factual Benchmarking (Level: Remember)

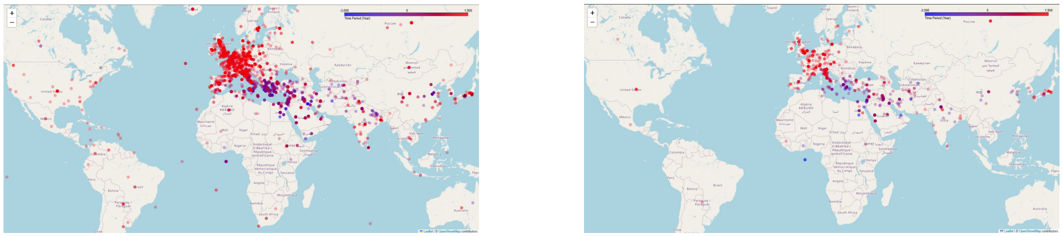
Each of the 1,007 selected events was input into GPT-4o Mini to generate three multiple-choice questions, corresponding to the fields of time, place, and main participants. These questions were designed to assess the model's factual recall and knowledge retention. Only events that were answered correctly by all models across these three questions were retained for higher-level benchmarking.

3.2.3 Higher-Order Benchmarking via Bloom's Taxonomy

To assess deeper historical reasoning beyond factual recall, we adopted Bloom's Taxonomy, a widely recognized framework for classifying educational learning objectives into six hierarchical cognitive levels (Anderson & Krathwohl, 2001). At the foundational level, *Remember* targets the retrieval of factual knowledge, such as dates, names, or specific events. The next level, *Understand*, involves grasping the meaning of historical content, such as summarizing a passage or interpreting a source. *Apply* requires learners to use historical knowledge in new contexts, for example, relating a past conflict to a contemporary situation. At a more advanced stage, *Analyze* focuses on breaking down historical narratives into components, identifying causes, effects, and relationships. The *Evaluate* level asks learners to make informed judgments, such as critiquing a historical decision or comparing the reliability of multiple sources. Finally, *Create* represents the highest cognitive level, involving the synthesis of new ideas or narratives based on historical understanding, such as constructing a counterfactual scenario or proposing an alternative interpretation of an event. This taxonomy informed the design of our evaluation framework, allowing us to probe different depths of reasoning, from simple recall to complex historical synthesis.

Question Generation Process: From the original set of 1,007 events, we first identified a subset of 394 events for which all tested models correctly answered the factual (i.e., ”Remember” level) questions. For each of these events, we then generated five multiple-choice questions, each aligned with one of the higher-order levels of Bloom's Taxonomy: *Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*. The initial versions of these questions were produced using the GPT-4o Mini model. Subsequently, the questions were refined and their cognitive complexity enhanced using the DeepSeek model to ensure greater depth and challenge across the higher taxonomy levels.

This structured approach enables a comprehensive evaluation of LLMs across both lower and higher order cognitive skills in the domain of historical reasoning.



(a) Geographic distribution of the full event dataset.

(b) Geographic distribution of the benchmarking subset.

Figure 3: Comparison of the geographic distributions in the full dataset and the benchmarking subset. Time is visualized using a gradient from blue (older events) to red (more recent events). The density of events in each geographic area is represented by color intensity, highlighting historically rich regions.

4 Methodology

4.1 Dataset Preparation and Preprocessing

We utilized digitized versions of three major historical texts: *The Story of Civilization* (Durant, 2016), *The History of al-Tabari* (al Tabari, 1967), and *Man's Religions*, the latter of which was digitized using Optical Character Recognition (OCR). Preprocessing involved cleaning the raw text and segmenting each book into smaller, coherent chunks. Each chunk was given a descriptive title and annotated with its start and end page numbers, based on a structural analysis of the text.

4.2 Event Extraction

We employed GPT-4 (32k context window) via prompt engineering to extract historical events from the preprocessed chunks. Two major challenges emerged in this process:

(i) Missing temporal and spatial information: In many cases, events lacked time or location data, both of which are essential for visualization on a temporal-spatial map. This issue stemmed either from limitations in the model's extraction capabilities or the absence of such details in the source text. To mitigate this, each prompt included both the target text segment and a set of recently extracted events to provide historical context. When time or place was not explicitly mentioned, the model was instructed to infer it based on its training data. A separate field was added to indicate whether this information was inferred (`True`) or directly stated (`False`).

(ii) Standardization of extracted fields: For consistency and usability, temporal data was converted into numeric formats (e.g., years, centuries), and spatial data into geographic coordinates (latitude and longitude). To support this, two additional fields were defined in the model prompt to extract standardized versions of time and location directly.

4.3 Evaluation of Extracted Events

To assess the quality of the extracted event dataset, a random sample of 50 events was selected for manual verification. Two independent evaluators reviewed each event's fields—including time, place, main participants, and others—labeling them as correct or incorrect based on careful examination of the original text and additional historical sources. Table 3 presents the results of this evaluation, including individual assessments and their average, demonstrating strong overall performance with an average accuracy of 94.1%. Notably, the standardization of place information exhibited slightly lower accuracy, reflecting challenges in precisely identifying geographical coordinates. These results indicate that the dataset is both robust and reliable for capturing critical historical event information.

Table 3: Evaluation of extracted events based on annotations by two dependent human annotators

	time	time estimated	time standard	place	place estimated	place standard	main participants	pages	total
annotator 1	90%	96%	98%	94%	100%	88%	96%	100%	95.25%
annotator 2	88%	96%	88%	94%	98%	88%	98%	94%	93%
average	89%	96%	93%	94%	99%	88%	97%	97%	94.125%

4.4 LLM Evaluation

We evaluated the performance of three large language models: GPT-4o Mini (OpenAI et al., 2024), DeepSeek-V3 (DeepSeek-AI et al., 2025), and Gemma-3-27B-IT (Team et al., 2025). Evaluation proceeded in two stages:

Factual Benchmarking (Remember level): Each model was assessed using three multiple-choice questions per event, targeting the fields of time, place, and main participants.

Higher-Order Reasoning Benchmarking: Events for which all three models answered correctly at the factual level were selected to generate more advanced questions. These were mapped to the upper levels of Bloom's Taxonomy (*Understand*, *Apply*, *Analyze*, *Evaluate*, and *Create*) to evaluate the models' deeper historical reasoning capabilities.

5 Results

For each multiple-choice question, the model's response was evaluated against the ground truth to determine its correctness. Overall accuracy was then calculated based on the proportion of correct responses. Table 4 presents the performance of the three models at the *Remember* level, while

Table 5 reports their results across the remaining five levels of Bloom's Taxonomy. They offer a detailed view of how different large language models perform across various dimensions of historical understanding. Below are several key insights drawn from the evaluation data:

Overall Performance Levels. **(1)** All models generally perform better on higher-order cognitive tasks (like *Evaluate* and *Create*) compared to the *Remember* and *Understand* levels. **(2)** *DeepSeek-V3* consistently outperforms *GPT4o-mini* and *Gemma-3* across nearly all categories and Bloom's levels, indicating stronger historical reasoning and comprehension capabilities.

Remember Level (Table 4). **(1)** Models excel in recognizing *Place* and *Main Participants*, with accuracy around 90% or above, while performance on *Time* is considerably lower (66.5%–75.9%). This suggests temporal understanding remains more challenging than spatial or entity recognition at the factual recall level. **(2)** *DeepSeek-V3* leads on all three *Remember* subcategories, pushing its total accuracy to 88.65%, about 5 percentage points higher than the other two models.

Higher-Order Cognitive Levels (Table 5). **(1)** Accuracy improves progressively from *Understand* (approximately 74–84%) to *Evaluate* and *Create* levels (approximately 79–92%), demonstrating that models can perform well on complex reasoning tasks when provided with structured historical data. **(2)** *DeepSeek-V3* again ranks highest across all five levels, exceeding 90% accuracy at *Evaluate* and *Create*, suggesting a better grasp of complex historical concepts and analysis. **(3)** *Gemma-3* trails behind *GPT4o-mini*, especially at the *Apply*, *Analyze*, *Evaluate*, and *Create* levels, indicating weaker performance in applying and synthesizing historical information.

Table 4: Model performance at the remember level, showing the number of correct answers alongside the corresponding accuracy percentages

model	**time**	**place**	**main participants**	**total**
gpt4o-mini	670 (66.534%)	911 (90.466%)	943 (93.644%)	83.55%
deepseek-v3	764 (75.868%)	955 (94.836%)	959 (95.233%)	88.65%
gemma-3-27b-it	704 (69.911%)	911 (90.466%)	916 (90.963%)	83.78%

Table 5: Model performance on higher-order levels (bloom's taxonomy), showing the number of correct answers alongside the corresponding accuracy percentages

model	**understand**	**apply**	**analyze**	**evaluate**	**create**
gpt4o-mini	298 (75.63 %)	327 (82.99 %)	348 (88.32 %)	357 (90.60 %)	349 (88.57 %)
deepseek-v3	332 (84.26 %)	335 (85.02 %)	351 (89.08 %)	364 (92.38 %)	362 (91.87 %)
gemma-3-27b-it	291 (73.85 %)	301 (76.39 %)	312 (79.18 %)	327 (82.99 %)	311 (78.93 %)

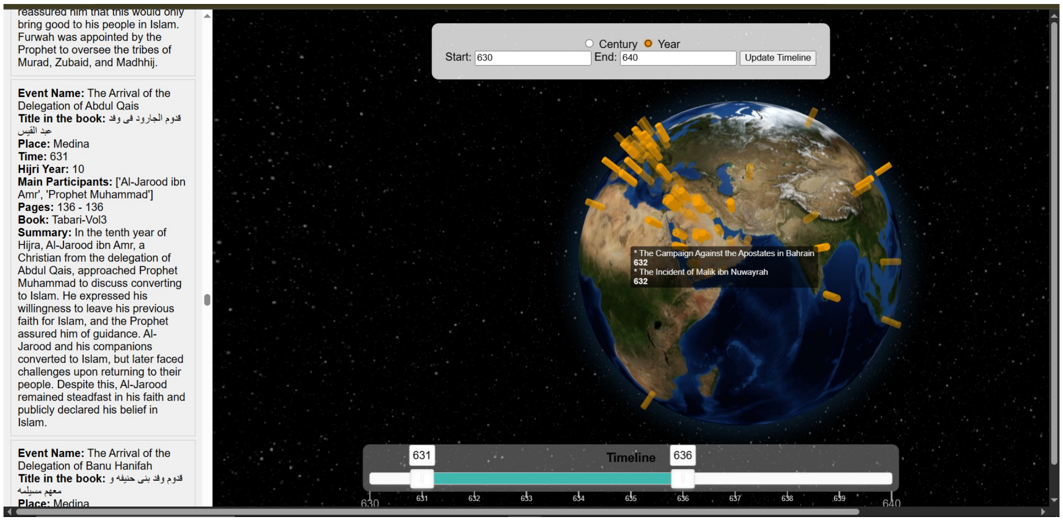
6 VISUALIZATION

To facilitate the exploration of the extracted historical events, we developed a web-based visualization platform featuring an interactive 3D globe. Users can select specific time intervals, by year or century, and view the corresponding events geographically displayed on the globe. An adjustable timeline is provided to further refine the temporal range and dynamically update the displayed events.

The intensity of the color bars on the map increases with the number of events associated with a given location; higher event density results in more saturated color markers. By hovering over a location, users can access a tooltip displaying detailed information about the associated events.

Additionally, a side panel presents a scrollable list of all currently filtered events, allowing for easier navigation and inspection.

This visualization platform is implemented using HTML and JavaScript, with the support of the Globe.GL library [2], a UI component built on Three.js/WebGL for interactive geographic data visualization. A screenshot of the interface is shown in Figure 4.


reassured him that this would only bring good to his people in Islam. Furwah was appointed by the Prophet to oversee the tribes of Murad, Zubaid, and Madhhij.

Event Name: The Arrival of the Delegation of Abdul Qais
Title in the book: قدوم الجارود فى وفد عبد القيس
Place: Medina
Time: 631
Hijri Year: 10
Main Participants: ['Al-Jarood ibn Amr', 'Prophet Muhammad']
Pages: 136 - 136
Book: Tabari-Vol3
Summary: In the tenth year of Hijra, Al-Jarood ibn Amr, a Christian from the delegation of Abdul Qais, approached Prophet Muhammad to discuss converting to Islam. He expressed his willingness to leave his previous faith for Islam, and the Prophet assured him of guidance. Al-Jarood and his companions converted to Islam, but later faced challenges upon returning to their people. Despite this, Al-Jarood remained steadfast in his faith and publicly declared his belief in Islam.

Event Name: The Arrival of the Delegation of Banu Hanifah
Title in the book: قدوم وفد بنى حنيفه و معهم مسيلمه
Place: Medina

○ Century ● Year
Start: 630 End: 640 Update Timeline

* The Campaign Against the Apostates in Bahrain
632
* The Incident of Malik ibn Nuwayrah
632

Figure 4: An example visualization of historical events on the interactive globe interface.

7 Conclusion

This paper introduced *HistoBench*, a comprehensive benchmark and dataset aimed at evaluating large language models' (LLMs) capabilities in understanding temporally grounded and context-rich historical narratives. By extracting and structuring over 13,000 events from diverse historical texts, we not only broadened the geographic and cultural scope of available historical datasets but also enabled meaningful analysis through an interactive globe-based visualization interface. Furthermore, we constructed a cognitively balanced benchmark of 1,007 multiple-choice questions, guided by Bloom's Taxonomy, to assess both factual recall and higher-order reasoning in history-focused tasks.

Our evaluation of three leading LLMs revealed notable performance differences across cognitive levels and question types, with DeepSeek-V3 demonstrating superior accuracy and reasoning consistency. These findings highlight both the potential and current limitations of LLMs in processing complex historical content. Further work may explore expanding the dataset to cover a broader range of cultures and historical traditions, as well as extracting additional layers of information, such as historical figures, their relationships, and interconnections, to enable more advanced forms of contextual and relational reasoning in historical language understanding.

[2] https://globe.gl/

REFERENCES

ACE. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*, 5.4.3 2005.07.01 edition, 2005.

A.J.M.J. al Tabari and F. Rosenthal. *The History of al-Tabari*. Bibliotheca Persica. State University of New York Press, 1988. ISBN 9780887065620. URL https://books.google.com/books?id=VEL8lWaqXtsC.

Ibn Jarir al Tabari. Tarikh al-rusul wa al-muluk (the history of al-tabari). https://www.ghbook.ir/index.php?option=com_dbook&task=viewbook&book_id=9678&lang=fa, 1967. Accessed online.

L.W. Anderson and D.R. Krathwohl. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, 2001. ISBN 9780801319037. URL https://books.google.com/books?id=EMQlAQAAIAAJ.

Harsimran Bedi, Sangameshwar Patil, Swapnil Hingmire, and Girish Palshikar. Event timeline generation from history textbooks. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Lung-Hao Lee, and Liang-Chih Yu (eds.), *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pp. 69–77, Taipei, Taiwan, December 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/W17-5912.

Nadav Borenstein, Natália da Silva Perez, and Isabelle Augenstein. Multilingual event extraction from historical newspaper adverts. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10304–10325, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.574. URL https://aclanthology.org/2023.acl-long.574.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models, 2024. URL https://arxiv.org/abs/2311.17667.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Xinya Du, Alexander Rush, and Claire Cardie. GRIT: Generative role-filler transformers for document-level event entity extraction. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 634–644, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.52. URL https://aclanthology.org/2021.eacl-main.52.

W. Durant. *The Story of Civilization*. The Story of Civilization. Simon and Schuster, 1942. URL https://books.google.com/books?id=T24gAAAAMAAJ.

Will Durant. The story of civilization (complete). https://archive.org/embed/TheStoryOfCivilizationcomplete, 2016. Accessed: 2016-12-22.

P. Fafalios, Yannis Marketakis, A. Axaridou, Yannis Tzitzikas, and M. Doerr. A workflow model for holistic data management and semantic interoperability in quantitative archival research. *Digital Scholarship in the Humanities*, 2023.

Giselle Gonzalez Garcia and Christian Weilbach. If the sources could talk: Evaluating large language models for research assistance in history, 2023. URL https://arxiv.org/abs/2310.10808.

Jakob Hauser, Daniel Kondor, Jenny Reddish, Majid Benam, Enrico Cioni, Federica Villa, James S. Bennett, Daniel Hoyer, Pieter Francois, Peter Turchin, and R. Maria del Rio-Chanona. Large language models'expert-level global history knowledge benchmark (hist-llm). In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 32336–32369. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/38cc5cba8e513547b96bc326e25610dc-Paper-Datasets_and_Benchmarks_Track.pdf.

Ashleigh Hawkins. Archives, linked data and the digital humanities: increasing access to digitised and born-digital archives via the semantic web. *Archival Science*, 2021.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Natalie Hervieux, Peiran Yao, Susan Brown, and Denilson Barbosa. Language resources from prominent born-digital humanities texts are still needed in the age of llms. *NLP4DH*, 2024.

Daniel Hienert and Francesco Luciano. Extraction of historical events from wikipedia. In *KNOW@LOD*, 2012. URL https://api.semanticscholar.org/CorpusID:28503128.

Hieke Huistra and Bram Mellink. Phrasing history: Selecting sources in digital repositories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 49(4):220–229, 2016. doi: 10.1080/01615440.2016.1205964.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge, 2023. URL https://arxiv.org/abs/2211.08411.

Amr Keleg and Walid Magdy. Dlama: A framework for curating culturally diverse facts for probing the knowledge of pretrained language models, 2023. URL https://arxiv.org/abs/2306.05076.

Viet Dac Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. Event extraction from historical texts: A new dataset for black rebellions. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2390–2400, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.211. URL https://aclanthology.org/2021.findings-acl.211.

Susan Leavy, Gerardine Meaney, Karen Wade, and Derek Greene. Curatr: A platform for semantic analysis and curation of historical literary texts. *ArXiv*, abs/2306.08020, 2019. URL https://api.semanticscholar.org/CorpusID:203165320.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In Kalina Bontcheva and Jingbo Zhu (eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-5010. URL https://aclanthology.org/P14-5010.

Thien Huu Nguyen and Ralph Grishman. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

John B. Noss. *Man's Religions*. The Macmillan Co., 1956.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. Gpt-4o system card, 2024. URL `https://arxiv.org/abs/2410.21276`.

Jiahao Qiu, Fulian Xiao, Yimin Wang, Yuchen Mao, Yijia Chen, Xinzhe Juan, Shu Zhang, Siran Wang, Xuan Qi, Tongcheng Zhang, Zixin Yao, Jiacheng Guo, Yifu Lu, Charles Argon, Jundi Cui, Daixin Chen, Junran Zhou, Shuyao Zhou, Zhanpeng Zhou, Ling Yang, Shilong Liu, Hongru Wang, Kaixuan Huang, Xun Jiang, Yuming Cao, Yue Chen, Yunfei Chen, Zhengyi Chen, Ruowei Dai, Mengqiu Deng, Jiye Fu, Yunting Gu, Zijie Guan, Zirui Huang, Xiaoyan Ji, Yumeng Jiang, Delong Kong, Haolong Li, Jiaqi Li, Ruipeng Li, Tianze Li, Zhuoran Li, Haixia Lian, Mengyue Lin, Xudong Liu, Jiayi Lu, Jinghan Lu, Wanyu Luo, Ziyue Luo, Zihao Pu, Zhi Qiao, Ruihuan Ren, Liang Wan, Ruixiang Wang, Tianhui Wang, Yang Wang, Zeyu Wang, Zihua Wang, Yujia Wu, Zhaoyi Wu, Hao Xin, Weiao Xing, Ruojun Xiong, Weijie Xu, Yao Shu, Yao Xiao, Xiaorui Yang, Yuchen Yang, Nan Yi, Jiadong Yu, Yangyuxuan Yu, Huiting Zeng, Danni Zhang, Yunjie Zhang, Zhaoyu Zhang, Zhiheng Zhang, Xiaofeng Zheng, Peirong Zhou, Linyan Zhong, Xiaoyin Zong, Ying Zhao, Zhenxin Chen, Lin Ding, Xiaoyu Gao, Bingbing Gong, Yichao Li, Yang Liao, Guang Ma, Tianyuan Ma, Xinrui Sun, Tianyi Wang, Han Xia, Ruobing Xian, Gen Ye, Tengfei Yu, Wentao Zhang, Yuxi Wang, Xi Gao, and Mengdi Wang. On path to multimodal historical reasoning: Histbench and histagent, 2025. URL `https://arxiv.org/abs/2505.20246`.

Marco Rovera, F. Nanni, and Simone Paolo Ponzetto. Providing advanced access to historical war memoirs through the identification of events, participants and roles. *arXiv.org*, 2019.

Rachele Sprugnoli and Sara Tonelli. Novel event detection and classification for historical texts. *Computational Linguistics*, 45(2):229–265, June 2019. doi: 10.1162/coli_a_00347. URL `https://aclanthology.org/J19-2002`.

Martina Katalin Szabó, Orsolya Ring, B. Nagy, L. Kiss, Júlia Koltai, Gábor Berend, László Vidács, László Vidács, A. Gulyás, and Zoltán Kmetty. Exploring the dynamic changes of key concepts of the hungarian socialist era with natural language processing methods. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 2020.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

J. F. Wakabayashi. Digital approaches to translation history. *The International Journal of Translation and Interpreting Research*, 2019. URL https://api.semanticscholar.org/CorpusID:201388968.

Yuqing Wang and Yun Zhao. Tram: Benchmarking temporal reasoning for large language models, 2024. URL https://arxiv.org/abs/2310.00835.

A Appendix

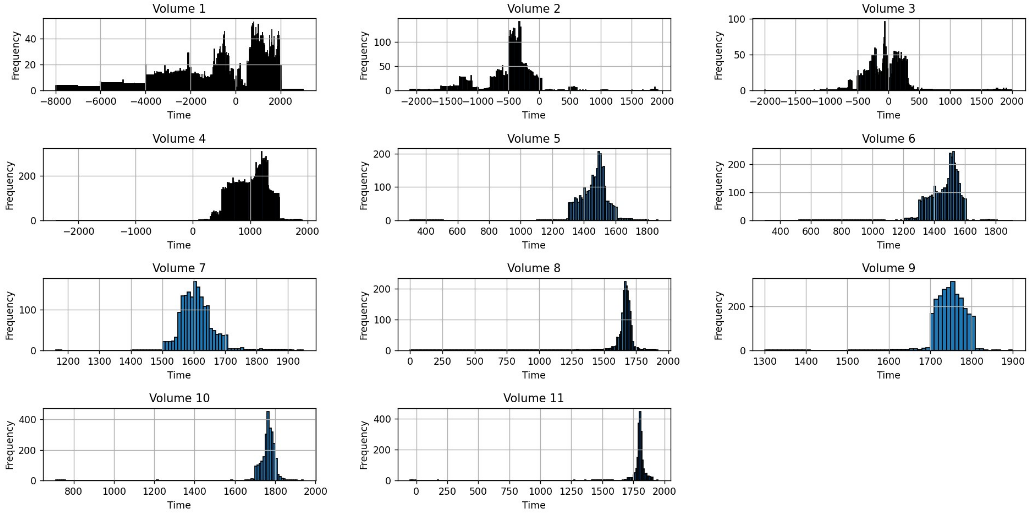


Figure 5: Bar chart showing the temporal distribution of extracted events by volume.