electronics

MDPI

Article

Knowledge Extraction from LLMs for Scalable Historical Data Annotation

Fabio Celli * and Dmitry Mingazov

Maggioli Research, Via Bornaccino 101, 47822 Santarcangelo di Romagna, Italy
* Correspondence: fabio.celli@maggioli.it

Abstract: This paper introduces a novel approach to extract knowledge from large language models and generate structured historical datasets. We investigate the feasibility and limitations of this technique by comparing the generated data against two human-annotated historical datasets spanning from 10,000 BC to 2000 CE. Our findings demonstrate that generative AI can successfully produce historical annotations for a wide range of variables, including political, economic, and social factors. However, the model's performance varies across different regions, influenced by factors such as data granularity, historical complexity, and model limitations. We highlight the importance of high-quality instructions and effective prompt engineering to mitigate issues like hallucinations and improve the accuracy of generated annotations. The successful application of this technique can significantly accelerate the development of reliable structured historical datasets, with a potentially high impact on comparative and computational history.

Keywords: large language models; knowledge extraction; cliodynamics

1. Introduction

Structured or semi-structured historical datasets, such as Seshat [1] and Chronos [2], are valuable tools for extracting models of societal evolution, cultural change [3], response to social crises [4] and geo-political patterns [5]. For example, data and models demonstrate that the evolution of warfare techniques alone can account for more than half of the increase in the hierarchical complexity of past societies. Additionally, improvements in governance specialization explain one-third of the increase in agricultural productivity [6].

Unfortunately, the development of trusted historical datasets is costly, time-consuming, not scalable and prone to errors and limitations. There are three open challenges in particular:

- Data integrity is a primary challenge in developing datasets of this nature, as certain phenomena, such as immaterial historical records, are difficult to quantify [7]. Specifically, historical data are inherently biased due to missing records [8], a problem that becomes more pronounced as we examine increasingly distant past periods.
- Subjectivity poses another significant challenge. Reaching consensus on historical data annotation is difficult due to the subjective nature of historical interpretation.
- A third challenge is knowledge representation, particularly how to compress historical information into a structured format. The Seshat dataset contains data about polities from 35 different natural geographic areas (NGAs) and provides many dimensions that report the presence or absence of a cultural trait in a polity at a specific point in time, for example, the presence/absence of copper, fortifications, firearms, written literature, coins and many others. These cultural dimensions can be computationally represented using one-hot encoding (OHE), a common technique for converting categorical features into numerical ones. OHE represents each category as a binary vector, where a '1' indicates presence and a '0' indicates absence. However, OHE often leads to high information sparsity in the transformed data, meaning that many values are zero. This sparsity can hinder the extraction of models using statistical or machine learning techniques [9].

Citation: Celli, F.; Mingazov, D. Knowledge Extraction from LLMs for Scalable Historical Data Annotation. *Electronics* **2024**, *13*, 4990. https://doi.org/10.3390/electronics13244990

Academic Editors: Lanting Fang and Yubo Song

Received: 14 November 2024
Revised: 14 December 2024
Accepted: 17 December 2024
Published: 18 December 2024

Time-series interpolation techniques are employed to address data integrity issues and crowdsourcing to address subjectivity. While they have a positive effect [10], crowdsourcing is ineffective in mitigating subjectivity, as human judgment is often influenced by cultural biases [11]. This work mainly addresses the challenge of knowledge representation and contributes to the improvement of AI-generated historical data with the application of a new compression technique.

1.1. Related Work

To solve the knowledge representation problem, one-hot encoded features are usually transformed into coarse-grained variables with compression techniques like principal component analysis (PCA) or autoencoders [12]; however, once compressed, the data lose their transparency, making it very difficult to interpret the models from a human perspective. However, techniques like time-resolved variables (TRVs) can compress historical data into meaningful scales [13]. In essence, TRVs transform categorical variables into interpretable numerical scales. Each level of the scale represents a developmental stage of a specific dimension, ordered chronologically by its first known historical appearance. For instance, a scale of military technologies might place stone at the first level, copper at the second, bronze at the third, iron at the fourth, and so on.

The hypothesis that motivates this work is that knowledge extraction from large language models (LLMs) can be exploited to annotate structured historical datasets in a scalable and cheaper way. The field of knowledge extraction from LLMs is quite novel, but there are successful applications. For example, prompts to perform knowledge extraction in structured data are used for testing the risk of LLMs to leak sensitive information [14]. Another successful application is tabular data augmentation (TDA), which enhances an existing dataset in a tabular format with additional structured data, with the aim of improving machine learning models [15]. However, the most popular use of LLMs for generating structured data might be synthetic data generation: systems like DiffLM, based on autoencoders and diffusion models [16], can generate tabular data similar to an original table, with the same data distribution but without privacy issues. A previous attempt to benchmark LLMs in historical data generation showed that LLMs can generate correct outputs, but there is considerable room for improvement in this task [17]. Crucially, recent advancements in LLMs have substantially mitigated their tendency to hallucinate, enabling the first real-world applications of LLM-based tabular data generation. In this domain, context-rich prompts have proven to significantly enhance both data generation quality and training efficiency [18].

This work is the first attempt to exploit knowledge extraction from LLMs for producing structured historical datasets in a scalable manner. To do so, we review evaluation methods in a previous study on the generation of structured data with LLMs (Section 1.2); then, we define our method (Section 2), provide description of the data Section 2.1, describe the experiments and the results of our evaluation (Section 3) and then draw our conclusion and outline potential future directions (Section 4).

1.2. Evaluation of Generative Models

It is not easy to evaluate the validity of AI-generated, synthetic structured data. The literature suggests five main metrics for this kind of evaluation [19]:

- The Pearson correlation coefficient is a general purpose measure that quantifies the linear relationship between two variables. It ranges from -1 to 1, with larger absolute values indicating a stronger linear correlation. This metric can quantitatively assess the similarity between generated and real data. When dealing with time series, like in historical data, the correlation coefficient can also evaluate how much the trend of generated and real data is similar [20].
- F1-score, the harmonic mean of precision and recall, is a metric to assess the accuracy of classification models. It measures how well a model's predicted discrete values

align with the ground truth values. This metric is often used in zero-shot or few-shot classification tasks [21].

- The Kolmogorov–Smirnov test, a statistical method used to determine if two datasets originate from the same underlying distribution, compares the cumulative distribution functions of two datasets and it is used in synthetic numerical data generation [22].
- Kullback–Leibler divergence is a metric that measures the difference between two probability distributions. A lower score signifies greater similarity between the two distributions. Unlike the Kolmogorov–Smirnov test, which uses statistics to express this difference, Kullback–Leibler divergence quantifies the information difference between one distribution and another [23].
- Edit distance, often referred to as Levenshtein distance, is a metric used to quantify the difference between two sequences and takes values from 0 to infinity. Applied on text, it measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to transform one string into another. It is used in the evaluation of code similarity [24].

We select average edit distance and correlation coefficient as metrics to evaluate structured historical data. Average edit distance is an error metric (lower scores mean greater similarity), while correlation coefficient should be close to 1 to indicate a good performance. Previous work in structured data generation obtained a minimum correlation of $\rho = 0.693$ on census income data [25] and a maximum $\rho = 0.983$ on air carrier statistics data [26], We refer to these results as baseline, as no previous evaluation is available for generated historical data.

2. Materials and Methods

Figure 1 illustrates our method for generating and evaluating structured historical data. We begin with Seshat and Chronos, two human-annotated historical datasets detailed in Section 2.1. First, we interpret the feature values for each of the two datasets in order to design specific instructions on how to categorize, extract and structure data. Then, we incorporate these instructions into the corresponding prompts (see Section 2.2).

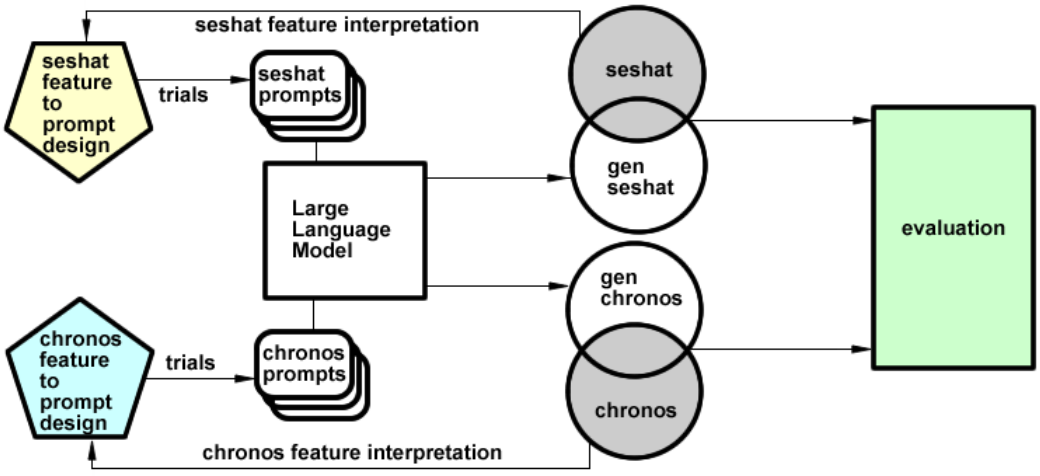

Figure 1. Schema of the method adopted.

In order to cover a wide spectrum of cases, we perform three different trials with Gemini 1.5 flash, Llama 3.1 and GPT-4o, testing parameters such as temperature and grounding, without any fine-tuning (see Section 3). We perform three runs with the same prompts, one per dataset, changing the NGAs: Turkey (which had a early development), Japan (intermediate development) and Ecuador (late development). The input to the LLM includes the prompt and an example of the desired output taken from the original data. The prompts contain instructions to estimate the desired categories, and there is the possibility to generate null values. Then, we align the generated data with the original dataset via timestamping.

Finally, we evaluate the correlation and edit the distance and coverage for each trial, run and dataset to obtain a comparison of the synthetic and original data.

2.1. Data

We use two publicly available datasets annotated with historical records: Seshat and Chronos. Seshat released many replication sets and we selected the social complexity dataset [27] which contains PCA-compressed variables about polities, such as population, territory, type of texts, type of government, currency, infrastructure, social scale and others, sampled from 30 NGAs with a timestep of 100 years. The weak point of Seshat is that the PCA-compressed variables are difficult to interpret and explain through instructions for the language model to replicate them. The strong point is that the timestep of 100 years is good for the alignment; hence, we expect high coverage. Polity identifiers in Seshat were not encoded following strict rules. Thus, we expect a poor edit distance in their replication.

Chronos is more detailed than Seshat, providing numerous variables for polities in 18 NGAs. These variables include quantitative data like population and territory size, as well as information expressed using TRVs. The use of TRVs simplifies interpretation, and this is a valuable feature of the Chronos dataset. Additionally, the consistent encoding of polity identifiers suggests potential for accurate alignment and low edit distances. However, Chronos's 10-year timestep might hinder alignment with generated datasets, possibly leading to lower coverage.

Gemini 1.5 flash, Llama 3.1 and GPT-4o have a knowledge cutoff between October and December 2023. Seshat has been published for the first time in 2017, so it is possible that LLMs contain it in a raw format. Chronos, instead, has been published in December 2024 and was not available at the time we ran the experiments; hence, it is impossible that LLMs contain it or access it with grounding.

2.2. Prompts

We create two specific prompts, one for Seshat and one for Chronos, which contain explanation for data annotation based on feature interpretation. We select a subset of comparable features from both datasets. For Seshat, the selected features are the following: **polity population** (PolPop) is the log-transformed polity population and **polity territory** (PolTerr) is the log-transformed territory size;

type of money (economy) is the OHE-PCA summarizing the presence/absence of precious metals, tokens, articles, paper currency, indigenous and foreign coins;

type of texts (infomedia) is the OHE-PCA of calendars, lists, sacred text, practical literature, scientific literature, fiction, and philosophy;

infrastructure (infrastructure) is the OHE-PCA of bridges, markets, mines and quarries, canals, roads, drinking water supply, ports, irrigation or production systems, food storage sites;

Social scale (scale) is the OHE-PCA of polity population, polity territory and the population of the largest settlement and **Specialization of governance** (government): the OHE-PCA of presence or absence of full-time bureaucrats, professional officers, priests, soldiers, examination systems, specialized buildings for government, full-time judges, formal legal code, courts, lawyers, merit promotion.

As part of the feature interpretation phase, all variables in Seshat were normalized to a 0–1 scale. This operation allowed us to manually check the descriptions of three polities at each scale level (0.1, 0.2, 0.3 and so on) to interpret these values and create textual instructions for the prompt. For instance, a population value of 0.1 roughly corresponded to a polity with 100–1000 people. All instructions are detailed in Figure 2. This method is inherently error-prone, and we expect a high edit distance. We also ask the LLM to generate a description of the historical period in order to perform a qualitative assessment of the output.

```
Generate a table in json format with information about the polities that exist in the [NGA] between 10000 BC and 2010 CE.
The structure of the table must be a century per row and the following columns:
NGA: location of the society or site;
polID: the polity ID, as the concat of two characters for civilization (where the population come from, use the two char state
abbreviation), three characters for culture (use abbreviation of population name, ruling dynasty or type site), one char for the
social type (c=community, n=nomad community, k=kingdom, e=empire, r=republic) and one character for the civilization stage (i=incipient,
f=formative, e=early, m=middle, l=late, t=terminal, *=any);
time:  (values below 0 represent years BC, timestep of 100 years, for example, -1000, -900, -800 and so on);
''PolPop'':  polity population:  0=around 100 people, 0.1=between 100 and 1000 people, 0.2=between 1000 and 5000 people, 0.3=between 5 and
10 thousand people, 0.4=between 10 and 50 thousand people, 0.5=between 50 and 100 thousand people, 0.6=between 100 thousand and 1 million
people, 0.7=between 1 and 100 million people, 0.8=between 100 and 500 million people, 0.9=between 500 million and 1 billion people,
1.0=population above 1 billion;
''PolTerr'':  polity territory class:  0.1=less than 100 km2, 0.2=between 100 and 500 km2, 0.3=between 500 and 1000 km2, 0.4=between 1
and 10 thousand km2, 0.5=between 10 and 50 thousand km2, 0.6=between 50 and 100 thousand km2, 0.7=between 100 and 500 thousand km2,
0.8=between 500 thousand and 1 million km2, 0.9=between 1 and 10 million km2, 1.0=more than 10 million km2;
government:  0=no institutions, 0.1=government by a chief, 0.2=government of a chief over other rulers, 0.3=government with ruler
and aristocrats, 0.4=government with an assembly, 0.5=government with assembly and religious institutions, 0.6=rule of more than
one religious institutions, 0.7=rule of one religious/ideological institution, 0.8=imperial rule of religion or ideology over other
governments, 0.9=imperial religious/ideological rule with bureaucracy, 1.0=government with religion/ideology bureaucracy and legal
institutions;
infrastructure:  level of infrastructures:  0=no infrastructures, 0.1=routes, 0.2=quarries, 0.3=monuments or special buildings,
0.4=production and irrigation systems, 0.5=urban markets, 0.6=canals dikes or aqueducts, 0.7=ports, 0.8=supply transport systems,
0.9=small portual systems, 1.0=large portual systems;
infomedia:  level of mediated information:  0=inferred oral tradition, 0.1=calendars, 0.4=symbols, 0.5=written laws, 0.7=presence of
religious/philosophical texts, 0.8=presence of written literature, 0.9=presence of written fiction, 1.0=presence of all kind of written
texts;
economy:  level of economy:  0=subsistence, 0.1=subsistence and barter, 0.2=barter, 0.3=use of tokens and barter, 0.5=use of tokens
and precious metals, 0.6=use of precious metals and foreign coins, 0.7=use of coins, 0.8=credit system, 0.9=banks, 1.0=paper currency,
1.1=stock market
scale:  level of social scale complexity as 0=hunter-gatherer groups, 0.1=nomad hunters and sparse villages, 0.2=sparse villages,
0.3=large villages, 0.4=simple domains, 0.5=complex domains, 0.6=kingdoms, 0.7=archaic state, 0.8=state, 0.9=state-empire, 1.0=empire,
1.1=state-nation
description:  text notes with a summary of maximum 400 characters about historical events happened in the century.  Focus on the
following types of events:  wars or battles; reforms; rulers; population; elites; disasters or epidemics; alliances or treaties;
socio-conomic context; famines or financial stress; protests or movements; changes of elite; main type site; religions and philosophies.
When possible, report the scientific references about the information in the text notes.
When values are unknown, set ''null'' to a value.
Below is an example of the desired output.

''NGA'': ''Konya Plain'', ''PolID'': ''TrBrzER'', ''Time'':  ''-2300'', ''PolPop'':  ''1.0'', ''PolTerr'':  ''1.0'', ''government'':  ''0.2'', ''infrastr'':  ''0.4'',
''infomedia'':  ''0.0'', ''economy'':  ''0.5'', ''scale'':  ''0.3'', ''description'':  ''This period begins with controversy, because the transition from
Late Chalcolithic to Early Bronze Age is not clear.  Some scholars argue that beginning of the Early Bronze Age should be dated to around
3000 BCE.''
```

Figure 2. Prompt for Seshat social complexity dataset generation.

In the Chronos dataset, all variables except population and territory were encoded as TRVs. This allowed for straightforward interpretation because, as reported in Section 1, the meaning of scales expressed as TRVs is known in advance.

The features selected from the Chronos dataset in order to be comparable to the ones in Seshat are the following: **polity population** (PolPop) is the estimated population in the polity at the time as an integer and **polity territory** (PolTerr) is the estimated size of polity territory in km^2.

Economic scale (economy) encodes the progression of economic evolution, with subsistence and exogamy [28] at the base level, bartering with precious goods at the first level [29], metals and weights at the second [30], coinage at the third, paper money as the fourth and stock market the fifth level [31].

Information management (infomedia) encodes the level of information and amount of contents that a culture can elaborate. The basic stage of this scale is oral tradition with its earliest evidence dating to 38,000 BC [32]. The first level in the scale is symbolism, which appeared around 9000 BC in Göbekli Tepe, Turkey [33], and around 6000 BC in Jiahu, China [34]. The second level is administration with written text, such as lists, laws or calendars [35]; at the third stage, there are religious/philosophical/scientific texts [36], followed by fiction at the fourth level [37] and news and opinions at the fifth level [38].

Infrastructure scale (infrastructure) encodes the level of control over infrastructures. The first stage includes routes and quarries [39]; at the second level, there are special buildings like circular towers in Tell Qaramel that are dated around 10,000 BC [40]. At the third stage, there are irrigation and production systems [41]: irrigation for the polities relied on massive agriculture and production systems for those polities who relied mainly on animal husbandry [42]. At the fourth step, there are urbanization and markets [43], followed by portual systems for goods' supply, evolved through the bronze, iron and early

modern ages [44]. At the sixth stage, there are railways and telecommunications, while the seventh, space systems, started in 1998 with the launch of the International Space Station.

Social scale (scale): encodes the evolutive steps of societies growing in population and territory [45]. At the base level, there are small groups of hunter–gatherers, dominant up to the Neolithic Period [46]; at the first level, there are sparse villages, started after the agricultural revolution [47]. At the second level, there are simple domains with a capital and connected villages, a dominant pattern in the chalcolithic [48]. At the third step, there are complex domains or archaic empires, started with Sargon of Akkad around 2500 BC [49]. At the fourth step, there are archaic states up to one million people [50]. At the fifth level, there are states and empires up to several million people; at the sixth step of the scale, there are state-nations, up to hundred million people; finally, at the seventh step, there are large nations, up to one billion people and more.

Political system (government) encodes the development of the power limits imposed to the rulers. At the base level, there is a sole ruler or chiefdom; hence, there is no limit to their power at all. At the first level, there are collective assemblies, possibly appeared in large agricultural villages [51]. At the second level, we have representatives of the population, which started in the republican Rome with the Tribunes of the Plebs [52]. The third stage involves legal impeachment, which empowers an institution like the parliament to nullify the ruler's authority without resorting to physical harm. The instructions included in the Chronos prompt (Figure 3) are derived from these descriptions of TRVs.

```
Generate a table in json format with information about the polities existed in the [NGA] between 10000 BC and 2010 CE.
The structure of the table must be a century per row and the following columns:
NGA: location of the society or site;
PolID: the polity ID, as the concat of two characters for civilization (where the population come from, use the two char state
abbreviation), three characters for culture (use abbreviation of population name, ruling dynasty or type site), one char for the
social type (c=community, n=nomad community, k=kingdom, e=empire, r=republic) and one character for the civilization stage (i=incipient,
f=formative, e=early, m=middle, l=late, t=terminal, *=any) ;
Time:  (values below 0 represent years BC, timestep of 100 years, for example -1000, -900, -800 and so on);
PolPop:  estimated population in the polity at the time;
PolTerr:  estimated size of polity territory in km2;
government:  scale of government/politic system; 0.0=no limits to ruler (monarchy), 0.1=assembly (aristocracy), 0.2=population
representatives (democracy), 0.3=rule of law with ruler check (legal impeachment)
infrastructure:  level of infrastructures:  0.1=routes/quarries, 0.2=common/special buildings, 0.3=production systems/irrigation,
0.4=urbanization/markets, 0.5=portual good supply systems, 0.6=railways/telecommunications, 0.7=space supply system
infomedia:  level of mediated information:  0.0=presence of oral tradition/gossip, 0.1=presence of written symbols, 0.2=presence of
laws/calendars, 0.3=presence of religious/philosophical texts, presence of 0.4=literature/fiction, presence of 0.5=newspapers/press,
economy:  level of economy:  0.0=subsistence/exogamy, 0.1=barter with goods/livestock, 0.2=precious metals by weight, 0.3=coins,
0.4=paper currency, 0.5=stock market
scale:  0.0=hunter-gatherers groups, 0.1=sparse villages, 0.2=simple domain, 0.3=complex domain or archaic empire, 0.4=archaic state or
kingdom, 0.5=modern state or empire, 0.6=state-nation, 0.7=modern nation
description:  text notes with a summary of maximum 400 characters about historical events happened in the century.  Focus on the
following types of events:  wars or battles; reforms; rulers; population; elites; disasters or epidemics; alliances or treaties;
socio-conomic context; famines or financial stress; protests or movements; changes of elite; main type site; religions and philosophies.
When possible, report the scientific references about the information in the text notes.
When values are unknown put null as value.
Below is an example of the desired output.

''NGA'': ''Adana'', ''PolId'':  ''TrOttem'', ''time'':  ''1600'', ''PolPop'':  ''27014772'', ''PolTerr'':  ''4930769'', ''government'':  ''0.2'', ''infrastr'':
''0.5'', ''infomedia'':  ''0.4'', ''economy'':  ''0.3'', ''scale'':  ''0.5'', ''description'':  ''Sultans Suleiman II and Ahmed II. Great Turkish War
in the 1690s:  the armies of the Holy League pushed the Ottomans back to the Balkans.  End of expansionist policies in Europe, series
of grand viziers from the Köprülü family.  Conquest of Cyprus in 1571 led to the naval defeat against the Holy League in the Battle of
Lepanto'',
```

Figure 3. Prompt for Chronos data generation.

3. Experiments and Results

We performed the experiments as zero-shot tasks, without any fine-tuning of LLMs. Experiments include three different trials per dataset, with three different settings, and three different runs on the selected NGAs: Turkey, Japan and Ecuador. We compare three different LLMs: Gemini-1.5-flash002, Llama 3.1 and GPT-4o. The experimental settings aim to test also temperature and grounding. The temperature parameter controls randomization in token selection and may affect the generation of correct labels and polity identifiers. A higher temperature can lead to more freedom in token selection. We tested the default value (1) and a more strict setting with a temperature of 0.5. The grounding parameter links the model output to verifiable sources of information through a Google search.

This technology connects Gemini with world knowledge and up-to-date information on the Internet. A dynamic retrieval configuration evaluates whether a prompt requires knowledge about real world or recent events and activates grounding accordingly. For example, a prompt like "Who won the latest F1 grand prix?" activates grounding search that helps the LLM generate correct answers. This feature is only available for the Gemini model and is useful in situations where accuracy and reliability are important. We test the importance of this option against the other LLMs that do not provide this feature.

Results, reported in Table 1, show several interesting things:

- Gemini yields the best coverage and has competitive correlation coefficients; Llama has the best edit distance and GPT the best correlation. The result of GPT is consistent with the results reported in multiple-choice question answering about historical data [17].
- Grounding has a positive impact on coverage: repeating the experiments using Gemini without the grounding option yields lower coverage (0.250 on Seshat and 0.492 on Chronos). The results of Gemini without grounding are in line with the other LLMs.
- Larger timesteps do not yield higher coverage: we expected higher coverage with Seshat, which has a timestep of 100 years, but this is true only with Gemini, not with the other LLMs.
- Temperature has no clear impact on the task: in particular, we expected better results in terms of edit distance on Chronos, where polity identifiers are more strict than in Seshat, but there is no evidence of this.
- The highest correlation coefficients are in general obtained with the Chronos dataset: this suggests a stronger relationship between the generated narratives and the ground truth data. This might be due to the higher granularity of the Chronos dataset and the transparency of TRVs, which allowed us to create better instructions in the prompt.

Table 1. Results on the different experimental settings: t1 = temperature 1, t0.5 = temperature 0.5. The best results are marked in bold.

Dataset	Settings	Avg Coverage	Avg Edit Distance	Avg Correlation
seshat	gemini1.5f-t1-grounding	**0.690**	1.737	0.804
seshat	gemini1.5f-t0.5-grounding	0.660	1.758	0.765
chronos	gemini1.5f-t1-grounding	0.667	1.583	0.847
chronos	gemini1.5f-t0.5-grounding	0.602	1.675	0.871
seshat	llama3.1-t1	0.248	1.589	0.505
seshat	llama3.1-t0.5	0.284	**1.524**	0.558
chronos	llama3.1-t1	0.484	1.951	0.705
chronos	llama3.1-t0.5	0.475	1.851	0.491
seshat	gpt4o-t1	0.242	1.533	0.806
seshat	gpt4o-t0.5	0.327	1.583	0.868
chronos	gpt4o-t1	0.366	1.785	0.884
chronos	gpt4o-t0.5	0.345	1.637	**0.902**

Overall, the best performing LLM is Gemini with grounding because it has the highest coverage. Tt the same time, it is competitive in terms of correlation coefficients and edit distance. We also report per-variable details of the best runs.

3.1. Detailed Variable Analysis

Detailed results of the trials with the best results are reported in Table 2. Coverage is consistent within a single run because it is referred to the NGA. These details reveal that Turkey has generally the lowest coverage, suggesting challenges in annotating historical data for a region with a long history. In particular, the LLM tends to annotate the polities in the far past with a timestep of 1000 years in order to keep informative content when information is scarce. Ecuador, on the other hand, has the highest coverage because the ground-truth data contained only few centuries, starting from 1500 CE. In fact, the impossibility to compute some correlation coefficients for this region is due to the poor availability

of annotated data. Nevertheless, all LLMs produced data since 10,000 BC for this region. Gemini produced annotations starting from 9000 BC, reporting that there is a lack of evidence for that period; Llama reported data from neighboring cultures starting from 9800 BC, and GPT started from 5000 BC generally referring to "early hunter–gatherer groups". We could evaluate only the validity of the labels after 1500 CE, as the gold standard contains data only for that period of time. The edit distance shows that the models struggle with the "polity" variable, suggesting difficulties in accurately annotating polity identifiers, as expected. Crucially, correlation coefficients are high, especially for infomedia and infrastructure. This is surprising, especially in Seshat, given that the instructions extracted from the OHE-PCA variables are error-prone.

Table 2. Details of the trials with the best results.

Dataset/Run	Variable	Coverage	Edit Distance	Correlation
gemini1.5f-t1-seshat-turkey	polity	0.254	4.321	
gemini1.5f-t1-seshat-turkey	economy	0.254	0.892	0.817
gemini1.5f-t1-seshat-turkey	infomedia	0.254	2.285	0.901
gemini1.5f-t1-seshat-turkey	polpop	0.254	0.821	0.772
gemini1.5f-t1-seshat-turkey	polterr	0.254	0.964	0.527
gemini1.5f-t1-seshat-turkey	infrastr	0.254	1.464	0.857
gemini1.5f-t1-seshat-turkey	gov	0.254	1.142	0.927
gemini1.5f-t1-seshat-turkey	scale	0.254	0.964	0.884
gemini1.5f-t1-seshat-japan	polity	0.818	6.055	
gemini1.5f-t1-seshat-japan	economy	0.818	0.999	0.547
gemini1.5f-t1-seshat-japan	infomedia	0.818	2.388	0.899
gemini1.5f-t1-seshat-japan	polpop	0.818	0.833	0.848
gemini1.5f-t1-seshat-japan	polterr	0.818	0.833	0.712
gemini1.5f-t1-seshat-japan	infrastr	0.818	1.0	0.903
gemini1.5f-t1-seshat-japan	gov	0.818	0.944	0.489
gemini1.5f-t1-seshat-japan	scale	0.818	0.777	0.791
gemini1.5f-t1-seshat-ecuador	polity	1.000	6.000	
gemini1.5f-t1-seshat-ecuador	economy	1.000	1.000	1.000
gemini1.5f-t1-seshat-ecuador	infomedia	1.000	2.000	
gemini1.5f-t1-seshat-ecuador	polpop	1.000	1.000	
gemini1.5f-t1-seshat-ecuador	polterr	1.000	1.000	
gemini1.5f-t1-seshat-ecuador	infrastr	1.000	1.000	
gemini1.5f-t1-seshat-ecuador	gov	1.000	2.000	
gemini1.5f-t1-seshat-ecuador	scale	1.000	1.000	1.000
llama3.1-t0.5-seshat-turkey	polity	0.036	3.750	
llama3.1-t0.5-seshat-turkey	economy	0.036	1.000	0.992
llama3.1-t0.5-seshat-turkey	infomedia	0.036	1.750	0.970
llama3.1-t0.5-seshat-turkey	polpop	0.036	0.750	0.973
llama3.1-t0.5-seshat-turkey	polterr	0.036	1.500	0.812
llama3.1-t0.5-seshat-turkey	infrastr	0.036	1.000	0.926
llama3.1-t0.5-seshat-turkey	gov	0.036	1.750	0.846
llama3.1-t0.5-seshat-turkey	scale	0.036	0.500	0.990
llama3.1-t0.5-seshat-japan	polity	0.318	4.000	
llama3.1-t0.5-seshat-japan	economy	0.318	1.143	0.573
llama3.1-t0.5-seshat-japan	infomedia	0.318	1.429	0.933
llama3.1-t0.5-seshat-japan	polpop	0.318	1.286	0.918
llama3.1-t0.5-seshat-japan	polterr	0.318	1.571	0.874
llama3.1-t0.5-seshat-japan	infrastr	0.318	0.857	0.944
llama3.1-t0.5-seshat-japan	gov	0.318	1.000	0.811
llama3.1-t0.5-seshat-japan	scale	0.318	1.286	0.928
llama3.1-t0.5-seshat-ecuador	polity	0.500	4.000	
llama3.1-t0.5-seshat-ecuador	economy	0.500	2.000	−1.000
llama3.1-t0.5-seshat-ecuador	infomedia	0.500	0.500	
llama3.1-t0.5-seshat-ecuador	polpop	0.500	1.000	

Table 2. *Cont.*

Dataset/Run	Variable	Coverage	Edit Distance	Correlation
llama3.1-t0.5-seshat-ecuador	polterr	0.500	1.000	−1.000
llama3.1-t0.5-seshat-ecuador	infrastr	0.500	1.000	
llama3.1-t0.5-seshat-ecuador	gov	0.500	1.500	
llama3.1-t0.5-seshat-ecuador	scale	0.500	1.000	−1.000
gpt4o-t0.5-chronos-turkey	polity	0.070	4.750	
gpt4o-t0.5-chronos-turkey	economy	0.070	0.000	1.000
gpt4o-t0.5-chronos-turkey	infomedia	0.070	0.500	0.949
gpt4o-t0.5-chronos-turkey	polpop	0.070	3.500	0.956
gpt4o-t0.5-chronos-turkey	polterr	0.070	3.750	0.989
gpt4o-t0.5-chronos-turkey	infrastr	0.070	0.500	0.949
gpt4o-t0.5-chronos-turkey	gov	0.070	1.000	0.187
gpt4o-t0.5-chronos-turkey	scale	0.070	0.750	0.944
gpt4o-t0.5-chronos-japan	polity	0.217	4.800	
gpt4o-t0.5-chronos-japan	economy	0.217	0.600	0.866
gpt4o-t0.5-chronos-japan	infomedia	0.217	0.600	0.986
gpt4o-t0.5-chronos-japan	polpop	0.217		
gpt4o-t0.5-chronos-japan	polterr	0.217		
gpt4o-t0.5-chronos-japan	infrastr	0.217	0.600	0.938
gpt4o-t0.5-chronos-japan	gov	0.217	0.800	0.963
gpt4o-t0.5-chronos-japan	scale	0.217	0.600	0.973
gpt4o-t0.5-chronos-ecuador	polity	0.750	5.667	
gpt4o-t0.5-chronos-ecuador	economy	0.750	0.667	1.000
gpt4o-t0.5-chronos-ecuador	infomedia	0.750	1.333	0.945
gpt4o-t0.5-chronos-ecuador	polpop	0.750		
gpt4o-t0.5-chronos-ecuador	polterr	0.750		
gpt4o-t0.5-chronos-ecuador	infrastr	0.750	0.667	0.945
gpt4o-t0.5-chronos-ecuador	gov	0.750	1.000	0.756
gpt4o-t0.5-chronos-ecuador	scale	0.750	0.667	0.990

3.2. Limitations

LLMs differ in the generation of data about polity population and polity territory expressed as integer numbers. Gemini does not generate annotations for these variables, while Llama always tries to generate estimates, and GPT can generate population and territory estimates only for recent polities, although no command for estimating values was given in the prompt. This means that bins and ranges are required in the instructions, and precise data cannot be generated. From this perspective, TRVs have a positive impact on prompt design. A manual qualitative analysis revealed a hallucination of Gemini, which placed the Edo period both in 1700 and in 1900, annotating the wrong type of government in 1900. This means that hallucinations are possible even with grounding.

While this study demonstrates the feasibility of generating structured historical samples using LLMs, it is critical to emphasize that these samples cannot be directly integrated into Seshat or Chronos without rigorous verification by domain experts.

4. Discussion and Conclusions

This paper presented the first attempt to automatically generate an annotation of historical data by means of LLMs. Trusted historical datasets are valuable tools for extracting models of societal evolution, geo-political patterns and response to social crises. The manual annotation of trusted historical data are costly and unscalable; hence, generative AI can help to produce scalable annotations, reducing costs. We compared the performance of an LLM generating annotations on two historical datasets–Seshat and Chronos–starting from 10,000 BC to 2000 CE with a timestep of 100 years. This study revealed that historical annotation generation is possible and promising; however , there are challenges. The quality and quantity of the training data for each region could influence the model's performance. A lack of high-quality historical data for certain regions, like Ecuador, might contribute to

lower accuracy and difficulty in the evaluation. Nevertheless, we demonstrate that it is possible to instruct an LLM to recognize broad categories of history, such as economy and social scale, even if the complexity of historical events and political structures in different regions can pose challenges for the model. Finally, despite the possibility to ground the LLM with search engines, hallucinations might happen. We suggest that our evaluation technique can help to spot these cases. However, further research is needed in prompt engineering in order to keep this problem under control.

We believe that there is still room for improvement for future research in this topic. Exploring advanced language models and training techniques, such as fine-tuning on specific historical datasets, can enhance accuracy. Moreover, incorporating more contextual information and examples in prompts, such as geographical, cultural, and political factors, can improve the model's ability to generate accurate and nuanced historical labels. By addressing these factors, future research can aim to improve the accuracy and coverage of automatic historical annotation, particularly for complex regions like Turkey and Ecuador.

Author Contributions: Conceptualization, methodology and data curation: F.C.; software and experiments: D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the European Commission, grant 101120657: European Lighthouse to Manifest Trustworthy and Green AI—ENFIELD.

Data Availability Statement: The Chronos dataset is available online as a Google sheet (https://docs.google.com/spreadsheets/d/1OW6CtmUudN3WTJ1VvWRZYZdTWVEjDJGns6Q8_I6EBwk/edit?usp=sharing, accessed on 18 November 2024) for collaborative open science. The Seshat project provides replication datasets available online at https://seshat-db.com/downloads_page/ (accessed on 18 November 2024).

Acknowledgments: This research employed data from the Seshat Databank (seshatdatabank.info, accessed on 18 November 2024) under Creative Commons Attribution Non-Commercial (CC By-NC SA) licensing.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LLMs	Large language models
TRVs	Time-resolved variables
OHE	One-hot encoding
PCA	Principal component analysis
NGA	Natural geographic area
TDA	Tabular data augmentation

References

1. Turchin, P.; Whitehouse, H.; François, P.; Hoyer, D.; Alves, A.; Baines, J.; Baker, D.; Bartokiak, M.; Bates, J.; Bennet, J.; et al. An introduction to Seshat: Global history databank. *J. Cogn. Hist.* **2020**, *5*, 115–123.
2. Celli, F.; Basile, V. History Repeats: Historical Phase Recognition from Short Texts. In Proceedings of the CLIC-it 2024, Pisa, Italy, 4–6 December 2024.
3. Richerson, P.J. A Dynamic Analysis of American Socio-Political History. A Review of Ages of Discord: A Structural Demographic Analysis of American History by Peter Turchin (Beresta Books, 2016). *Cliodynamics* **2017**, *8*, 229–239.
4. Turchin, P. Multipath forecasting: The aftermath of the 2020 American crisis. In *How Worlds Collapse*; Routledge: Oxfordshire, UK, 2023; pp. 397–416.
5. Collins, R. A dynamic theory of battle victory and defeat. *Cliodynamics* **2010**, *1*, 3–25.
6. Turchin, P.; Whitehouse, H.; Gavrilets, S.; Hoyer, D.; François, P.; Bennett, J.S.; Feeney, K.C.; Peregrine, P.; Feinman, G.; Korotayev, A.; et al. Disentangling the evolutionary drivers of social complexity: A comprehensive test of hypotheses. *Sci. Adv.* **2022**, *8*, eabn3517.
7. Horsley, N. What can a knowledge complexity approach reveal about big data and archival practice? In Proceedings of the 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, USA, 11–14 December 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2246–2250.

8. Demartini, G.; Roitero, K.; Mizzaro, S. Data bias management. *Commun. ACM* **2023**, *67*, 28–32.
9. Ul Haq, I.; Gondal, I.; Vamplew, P.; Brown, S. Categorical features transformation with compact one-hot encoder for fraud detection in distributed environment. In *Data Mining: 16th Australasian Conference, AusDM 2018, Bahrurst, NSW, Australia, 28–30 November 2018*; Revised Selected Papers 16; Springer: Berlin/Heidelberg, Germany, 2019; pp. 69–80.
10. Oh, C.; Han, S.; Jeong, J. Time-series data augmentation based on interpolation. *Procedia Comput. Sci.* **2020**, *175*, 64–71.
11. Barbera, D.L.; Maddalena, E.; Soprano, M.; Roitero, K.; Demartini, G.; Ceolin, D.; Spina, D.; Mizzaro, S. Crowdsourced Fact-checking: Does It Actually Work? *Inf. Process. Manag.* **2024**, *61*, 103792.
12. Chiarot, G.; Silvestri, C. Time series compression survey. *ACM Comput. Surv.* **2023**, *55*, 1–32.
13. Celli, F.; Lepri, B. Feature Engineering for Quantitative Analysis of Cultural Evolution. *SocArXiv* **2023**. https://doi.org/10.31235/osf.io/aj8xk.
14. Yan, B.; Li, K.; Xu, M.; Dong, Y.; Zhang, Y.; Ren, Z.; Cheng, X. On protecting the data privacy of large language models (llms): A survey. *arXiv* **2024**, arXiv:2403.05156.
15. Cui, L.; Li, H.; Chen, K.; Shou, L.; Chen, G. Tabular data augmentation for machine learning: Progress and prospects of embracing generative ai. *arXiv* **2024**, arXiv:2407.21523.
16. Zhou, Y.; Wang, X.; Niu, Y.; Shen, Y.; Tang, L.; Chen, F.; He, B.; Sun, L.; Wen, L. DiffLM: Controllable Synthetic Data Generation via Diffusion Language Models. *arXiv* **2024**, arXiv:2411.03250.
17. Hauser, J.; Kondor, D.; Reddish, J.; Benam, M.; Cioni, E.; Villa, F.; Bennett, J.S.; Hoyer, D.; Francois, P.; Turchin, P.; et al. Large Language Models' Expert-level Global History Knowledge Benchmark (HiST-LLM). In Proceedings of the Thirty-Eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, Vancouver, BC, Canada, 10–15 December 2024.
18. Banday, B.; Thopalli, K.; Islam, T.Z.; Thiagarajan, J.J. On The Role of Prompt Construction In Enhancing Efficacy and Efficiency of LLM-Based Tabular Data Generation. *arXiv* **2024**, arXiv:2409.03946.
19. Kim, D.K.; Ryu, D.; Lee, Y.; Choi, D.H. Generative models for tabular data: A review. *J. Mech. Sci. Technol.* **2024**, *38*, 4989–5005.
20. Lin, Y.T.; Chen, Y.N. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv* **2023**, arXiv:2305.13711.
21. Abburi, H.; Suesserman, M.; Pudota, N.; Veeramani, B.; Bowen, E.; Bhattacharya, S. Generative ai text classification using ensemble llm approaches. *arXiv* **2023**, arXiv:2309.07755.
22. Siska, C.; Marazopoulou, K.; Ailem, M.; Bono, J. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 10406–10421.
23. Wu, T.; Tao, C.; Wang, J.; Yang, R.; Zhao, Z.; Wong, N. Rethinking kullback-leibler divergence in knowledge distillation for large language models. *arXiv* **2024**, arXiv:2404.02657.
24. Song, Y.; Lothritz, C.; Tang, D.; Bissyandé, T.F.; Klein, J. Revisiting Code Similarity Evaluation with Abstract Syntax Tree Edit Distance. *arXiv* **2024**, arXiv:2404.08817.
25. Xia, J.; Zhang, S.; Cai, G.; Li, L.; Pan, Q.; Yan, J.; Ning, G. Adjusted weight voting algorithm for random forests in handling missing values. *Pattern Recognit.* **2017**, *69*, 52–60.
26. Valdiviezo, H.C.; Van Aelst, S. Tree-based prediction on incomplete data using imputation or surrogate decisions. *Inf. Sci.* **2015**, *311*, 163–181.
27. Turchin, P.; Brennan, R.; Currie, T.; Feeney, K.; Francois, P.; Hoyer, D.; Manning, J.; Marciniak, A.; Mullins, D.; Palmisano, A.; et al. Seshat: The global history databank. *Cliodynamics* **2015**, *6*, 77–107.
28. Dow, G.K.; Reed, C.G.; Woodcock, S. The Economics of Exogamous Marriage in Small-scale Societies. *Econ. Inq.* **2016**, *54*, 1805–1823.
29. Khalaily, H.; Valla, F.R. Obsidian in Natufian context: The case of Eynan (Ain Mallaha), Israel. In *The Natufian Foragers in the Levant. Terminal Pleistocene Social Changes in Western Asia*; Berghahn Books: Oxford, UK, 2013; pp. 193–202.
30. Ialongo, N.; Hermann, R.; Rahmstorf, L. Bronze Age weight systems as a measure of market integration in Western Eurasia. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2105873118.
31. Hafer, R.W.; Hein, S.E. *The Stock Market*; Bloomsbury Publishing USA: New York, NY, USA, 2006.
32. Matchan, E.L.; Phillips, D.; Jourdan, F.; Oostingh, K. Early human occupation of southeastern Australia: New insights from 40Ar/39Ar dating of young volcanoes. *Geology* **2020**, *48*, 390–394.
33. Ayaz, O. Self-Revelation: An Origin Myth Interpretation of the Göbekli Tepe Culture (An Alternative Perspective on Anthropomorphic Themes). *Yüzüncü Yıl Üniv. Sos. Bilim. Enstitüsü Derg.* **2023**, *60*, 191–208.
34. Li, X.; Harbottle, G.; Zhang, J.; Wang, C. The earliest writing? Sign use in the seventh millennium BC at Jiahu, Henan Province, China. *Antiquity* **2003**, *77*, 31–44.
35. Woods, C. The earliest Mesopotamian writing. In *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*; Institute for the Study of Ancient Cultures Museum: Chicago, IL, USA, 2010; pp. 33–50.
36. Willard, R.H. Weights and Measures in Egypt. In *Encyclopaedia of the History of Science, Technology, and Medicine in Non-Western Cultures*; Springer: Dordrecht, The Netherlands, 2008; pp. 2244–2251. https://doi.org/10.1007/978-1-4020-4425-0_8933.
37. Tigay, J.H. *The Evolution of the Gilgamesh Epic*; Bolchazy-Carducci Publishers: Wauconda, IL, USA, 2002.
38. Barker, H. *Newspapers and English Society 1695–1855*; Routledge: Oxfordshire, UK, 2014.

39. Aurenche, O.; Galet, P.; Régagnon-Caroline, E.; Évin, J. Proto-Neolithic and Neolithic cultures in the Middle East—The birth of agriculture, livestock raising, and ceramics: A calibrated 14C chronology 12,500–5500 cal BC. *Radiocarbon* **2001**, *43*, 1191–1202.
40. Mazurowski, R.F.; Michczyńska, D.J.; Pazdur, A.; Piotrowska, N. Chronology of the early Pre-Pottery Neolithic settlement Tell Qaramel, northern Syria, in the light of radiocarbon dating. *Radiocarbon* **2009**, *51*, 771–781.
41. Kurt, A. *Ancient near East V1*; Routledge: Oxfordshire, UK, 1996.
42. Makarewicz, C.A.; Arbuckle, B.S.; Öztan, A. Vertical transhumance of sheep and goats identified by intra-tooth sequential carbon (δ13C) and oxygen (δ18O) isotopic analyses: evidence from Chalcolithic Köşk Höyük, central Turkey. *J. Archaeol. Sci.* **2017**, *86*, 68–80.
43. Altenmüller, H. Old Kingdom: Fifth Dynasty. *Oxf. Encycl. Anc. Egypt* **2001**, *2*, 601.
44. Knapp, A.B. Thalassocracies in Bronze Age eastern Mediterranean trade: Making and breaking a myth. *World Archaeol.* **1993**, *24*, 332–347.
45. Turchin, P. *Ultrasociety: How 10,000 Years of War Made Humans the Greatest Cooperators on Earth*; Beresta Books: Chaplin, CT, USA, 2016.
46. Marlowe, F.W. Hunter-gatherers and human evolution. *Evol. Anthropol. Issues News Rev.* **2005**, *14*, 54–67.
47. Barker, G. *The Agricultural Revolution in Prehistory: Why Did Foragers Become Farmers?*; Oxford University Press: Oxford, UK, 2006.
48. Heyd, V. Growth and expansion: Social, economic and ideological structures in the European Chalcolithic. In *Is there a British Chalcolithic*; Oxbow Books: Oxford, UK, 2012; pp. 96–112.
49. Steinkeller, P. The Sargonic and Ur III Empires. In *The Oxford World History of Empire: Volume Two: The History of Empires*; Oxford University Press: Oxford, UK, 2020; Volume 43, pp. 43–72.
50. Adams, R.M. Complexity in archaic states. *J. Anthropol. Archaeol.* **2001**, *20*, 345–360.
51. Hodder, I. Staying egalitarian and the origins of agriculture in the Middle East. *Camb. Archaeol. J.* **2022**, *32*, 619–642.
52. Smith, C. The Origins of the Tribunate of the Plebs. *Antichthon* **2012**, *46*, 101–125.