Large Language Models Rival Human Performance in Historical Labeling

Fabio Celli[1][0000−0002−7309−5886] and Valerio Basile[2][0000−0001−8110−6832]

[1] Gruppo Maggioli, Santarcangelo di Romagna, Italy.
[2] University of Turin. Turin, Italy.
fabio.celli@maggioli.it

Abstract. This study examines the application of Large Language Models to automatically annotate the phases of the Structural-Demographic Theory from short descriptions of historical decades. This task is useful for understanding social instability, but it is inherently subjective and challenging due to the temporal nature of labels. A single misalignment in phase labeling between annotators can cascade through subsequent timesteps, causing the inter-annotator agreement to decrease exponentially. Our results indicate that models with more than 400 billion parameters achieve very high agreement, while models with fewer than 100 billion parameters are prone to hallucinations. Moreover, the two largest models we tested (GPT4 and Lama3.1-405b) reach inter-annotator agreement comparable to pairs of human annotators, paving the way towards automated annotation. However, the need for very large models could hinder the democratization of automatic historical annotation due to the required computational resources. To mitigate this, we suggest collaborations between universities and companies, in order to share knowledge and computational power.

Keywords: Large Language Models · Structural Demographic Theory · Historical Phase Labeling.

1 Introduction And Background

The Structural Demographic Theory (SDT) is a framework to understand and explain long-term social pressures that can lead to major outbreaks of sociopolitical instability, such as revolutions and civil wars [6]. This theory describes how demographic changes, economic inequalities, and the dynamics within the elite class cyclically create or alleviate pressures for instability [16] [7]. In fact, when combined with data modeling, SDT allowed researchers to accurately predict the global crises of the 2020s [17]. It also proved to be useful to analyze many historical events, such as the French Revolution's causes; the American Civil War's elite rivalries [18], and the Qing Dynasty's collapse [10]. Moreover, researchers successfully used the SDT framework to understand contemporary outbreaks of instability, from the Egyptian revolution of 2011 [9] to the US political instability in 2021 [19]. This wide record of applications demonstrates the

value of SDT for analyzing complex socio-political patterns in historical data [20], and recent research explored the possibility to annotate data with SDT. In particular, the SDT defines five cyclical historical phases: growth, population immiseration, elite overproduction, state stress and crisis. Despite it is possible to annotate SDT phases on textual data [1], the task proved to be very challenging. Crucially, the subjective nature of historical phase interpretation introduces significant inter-annotator variability, potentially influenced by personal bias, including Eurocentrism. While certain historical periods exhibit clear consensus (e.g., the French Revolution is a crisis), others, such as the French intervention in Mexico or the early Maoist era in China, reveal significant interpretive divergence, demonstrating the difficulty in establishing consistent phase boundaries. The predominant methodologies in historical data annotation involve the development of schemas and guidelines for event detection and categorization that are grounded in linguistic theory [14] or thesauri [12]. Literature on the application of these methodologies reports a fairy high Cohen's k [3], around 0.7, with twenty labels and two annotators. In contrast, the annotation of SDT phases from short texts exhibits limited inter-annotator agreement. The reported inter-annotator agreement for identifying SDT phases from brief historical descriptions is a poor Fleiss' k [4] of 0.206 [1]. Achieving high agreement on this specific task is inherently challenging due to the temporal nature of SDT phases, which introduces extra constraints than a normal annotation task. In practice, a single misalignment in phase labeling between two annotators can cascade through subsequent time-steps, causing the inter-annotator agreement to decrease exponentially. A fair agreement (Fleiss' k 0.455) can be achieved by first assigning a default SDT sequence (e.g., two decades of growth, two of population impoverishment, two of elite overproduction, three of state stress, and one of crisis) and then having annotators assess and modify these labels for each decade [1]. However, we suggest that this approach artificially inflates agreement. This is because the fewer changes annotators make to the initial sequence, the higher their resulting agreement will be. For this reason, we will use k=0.206 as reference baseline. State-of-the-art methods, such as the use of generative AI, proved to be successful in supporting annotation workflows [15], hence we propose the use of Large Language Models (LLMs) as annotators to address SDT phase labeling. LLMs-as-Annotators proved to be promising [11], and literature shows that they possess historical knowledge that can be leveraged [2]. In order to exploit LLMs-as-annotators for this task, we turn existing guidelines on SDT annotation into a prompt, then we execute this prompt with different LLMs annotating the same data, then we compute the inter- and intra-annotator agreement. Finally, we compare the results against a human baseline. The paper is structured as follows: in Section 2 we describe the SDT labels and the annotation schema, in Section 3 we describe the data we use, in Section 4 we report the prompt and the results of the annotation evaluation, and finally in Section 5 we draw our conclusions.

2 Annotation Schema

The SDT posits that long-term societal instability and upheaval are driven by cyclical interactions between demographic, economic, and political factors. These cyclical interactions, called secular cycles [21], are typically 75-100 years in duration [8], and are characterized by the interplay of three core actors through five distinct phases. The actors are the population, elites and the State. Population constitutes approximately 90% of society, the population functions as the primary source of labor and resources, with limited consumption of generated wealth. Elites represent roughly 8% of society and are responsible for problem-solving and constitute the pool of potential State members. Elite composition and mobility are contingent upon the prevailing governance structure. State is approximately 2% of society. It enforces governance and manages resource allocation. It consists of one or more elite factions, and serves to codify and perpetuate societal culture. These actors engage in five sequential phases, marked by increasing socio-political instability and defined as follows:

- Growth Phase: characterized by robust cultural cohesion, increased state control and trade network expansion, this phase leads societies towards stability, albeit with sustainability concerns. Post-war reconstruction periods exemplify this phase.
- Population Immiseration Phase: population growth outpaces economic expansion, driven by the disparity between capital return and wage growth [13]. This results in heightened inequality and social unrest, as the state's extractive capacity reaches its limits.
- Elite Overproduction Phase: increased population access to elite ranks strains social mobility mechanisms, diminishing elite problem-solving capacity and increasing societal instability.
- State Stress Phase: State governance and elite-population cooperation deteriorate, leading to elite fragmentation and potential civil conflict. The financial instability of the state makes it vulnerable to destabilizing events.
- Crisis, Collapse, or Recovery Phase: the state undergoes reformation or overthrow, culminating in a new social equilibrium and the initiation of a subsequent cycle.

This schema describes a process. However, in order to operationalize the guidelines into a prompt (we will call it "annotation prompt"), we need to find recognizable cues from text associated to the different labels, and turn them into clear instructions. To do so, we use another prompt (we will call it "knowledge extraction prompt"), and generate instructions from data, as described in the next section.

3 Data and Prompts

For the sake of reproducibility, and to allow comparison with previous studies, we opted to use the Chronos dataset [1]. Chronos is a historical dataset containing short, decade-by-decade textual descriptions of 366 polities sampled from 18

sampling points equally distributed around the world, and spanning from prehistory to the 2010s. It was compiled using Seshat [22] and Wikipedia, focusing on the collection of key historical events (wars, rulers, reforms, etc.) summarized within a 400-character limit. Chronos is manually annotated with SDT phases by an official human annotator, and evaluated by other two human raters on 93 examples, and 5 labels. Exploiting the existing SDT annotation, we extracted

Summarize the following description/phase pairs into prompt instructions optimized to classify the phases and reducing potential overlaps: ⟨*data*⟩

Fig. 1. Knowledge Extraction Prompt for the generation of instructions related to SDT labels. Executed with GPT-4.

60 examples of historical decades descriptions per label (for a total of 300 examples) and summarized them with the knowledge generation prompt [5] reported in Figure 1. In this way we obtained clear instructions that we included in the annotation prompt, reported in Figure 2.

The data we use in our experiments are the same as those used for evaluating the annotation of the Chronos dataset. It comprises 93 decades from 6 polities from different places and times: the Mexican Republic, the Later Jin Dynasty, the Mongol Empire, the Yuan Dynasty, the Roman Kingdom and the Early Roman Republic. The dataset is freely available for replication studies[3]. Examples follows:

- Decade: 380s b.C., polity: Early Roman Republic, description: The Romans were defeated by Brennus of Senones in the Battle of Allia River in 387 BC. The Senones besieged Rome but probably the health conditions were bad and accepted a ransom in gold and silver to leave the city. Brennus cheated when weighting the gold and Romans, helped by the returned Furius Camillus, defeated Brennus.
- Decade: 1170s, polity: Later Jin Dynasty, description: Emperor Shizong (r. 1161-1189) confiscated unused land from Jurchen landowners and redistributed to Jurchen farmers, but they preferred to lease the work to Chinese farmers and engage in drinking instead. in 1175 paper factory in Hangzhou employed more than a thousand Chinese workers. Integration problems (language and customs).
- Decade: 1290s, polity: Yuan dynasty, description: Kublai Kahn promoted commercial, scientific, and cultural growth. He supported the merchants of the Silk Road trade network by protecting the Mongol postal system. he also cancelled the Confucian exhamination and managed power in autocratic way.

[3] `https://huggingface.co/datasets/facells/chronos-llm-sdt-agreeement`

Act as an expert historian and consider the Structural Demographic Theory (SDT). Given a set of descriptions of historical decades for different polities, label each description with one of the following secular cycle phases (sdtphase):
Start of knowledge-generated prompt
0=crisis (in this phase may happen societal collapse patterns, power transitions, conflicts, administrative or social structure changes, and external influences. Look for signs of civil wars, military coups, environmental factors, population movements, reform of tax systems, trade network disruptions, class conflicts, and foreign invasions).
1=growth (a society recovers from a crisis finding a new fresh culture that creates social cohesion. to recognize this phase examine the power structure patterns, legitimacy of rule, social organization, cultural elements, military aspects, and social changes. Look for the presence of strong elite classes, religious legitimation of power, centralized administrative systems, trade networks, cultural practices, territorial expansion, and population movements);
2=population impoverishment (growth slows and inequalities begin to emerge. to recognize this phase evaluate the power dynamics, economic patterns, military aspects, cultural/religious elements, administrative features, and infrastructure development. Look for succession struggles, trade route development, territorial conquests, religious tolerance, bureaucratic reforms, and construction projects);
3=elite overproduction (the number elite aspirants rises and the social lift mechanisms deteriorate. To recognize this phase assess power dynamics, governance, economic patterns, social structures, cultural and technological development, and common catalysts for change. Look for power struggles, trade system developments, social unrest between elite and population, religious developments, and military conflicts),
4=state stress (elites struggle to institutionalize their advantages. to recognize this phase review political instability, power struggles, economic challenges, military conflicts, administrative changes, and social/religious tensions. Look for succession disputes, financial crises, territorial loss, reforms to advantage specific elite groups, social unrest and religious conflicts).
End of knowledge-generated prompt
A cycle cannot turn back and cannot skip phases. So if in 1940 there is a phase 0, in 1950 there should be a phase 1, in 1960 there can be a phase 1 or phase 2. If in 1960 there is a phase 2, in 1970 there can be a phase 2 or phase 3, not a phase 4. If in 1970 there is a phase 3, in 1980 there can be a phase 3 or 4, and if in 2000 there is phase 4, in 2010 there can be a phase 0 or another phase 4. The decade after phase 0 the cycle restarts from phase 1.
This is an example of the input (json): $\langle example \rangle$
and this is the desired output (csv): $\langle example \rangle$
set of descriptions to label (json): $\langle data \rangle$

Fig. 2. Annotation Prompt. The part generated with the Knowledge Extraction prompt is marked by tags. The secular cycle phases are defined as numeric labels: 1=growth, 2=population immiseration, 3=elite overproduction, 4=state stress, 0=crisis.

4 Experiments

We performed three experiments: 1) to test the feasibility of the task; 2) to compare the inter-annotator agreement of humans and LLMs; 3) to test intra-annotator agreement of LLMs. We used LLMs of different size: Mistral-7b (7.25 billion parameters); Llama3.3-70b (70 billion parameters); Mistral-Large-2411 (123 billion parameters); Llama-3.1-405b (405 billion parameters) and GPT-4 (1.8 trillion parameters). We set the temperature to zero in order to perform a deterministic inference and enhance reproducibility.

Inter-Annotator Agreement	**raters**	**examples**	**Fleiss' K**	**Cohen's K**
human baseline	3	93	0.206	-
all LLMs	3	93	*0.133*	-
mistral-large-2411 + gpt4	2	93	*0.081*	*0.095*
mistral-large-2411 + llama3.1-405b	2	93	*0.049*	*0.076*
llama3.1-405b + gpt4	2	93	**0.255**	**0.253**
human-official + human1	2	93	*0.138*	*0.139*
human-official + human2	2	93	0.232	0.234
human1 + human2	2	93	0.248	0.250
llama3.1 + official-human	2	93	0.206	0.206
llama3.1 + human1	2	93	**0.454**	**0.455**
llama3.1 + human2	2	93	0.262	0.263
gpt4 + official-human	2	93	0.211	0.214
gpt4 + human1	2	93	*0.104*	*0.109*
gpt4 + human2	2	93	0.278	0.280
Intra-Annotator Agreement	**trials**	**examples**	**Fleiss' K**	**Cohen's K**
llama3.1-405b	2	93	**0.641**	**0.642**
llama3.1-405b	3	93	0.572	-
gpt4	2	93	0.525	0.525
gpt4	3	93	0.331	-
human-official	2	93	0.519	0.510

Table 1. Evaluation of the Inter- and Intra-Annotator Agreement between LLMs and humans. The best result is marked in bold, the ones below the baseline are marked in italics.

The goal of experiment 1 is to correctly generate SDT labels formatted in csv. Some LLMs had input limitations, so we had to input one polity at a time. However, the smallest LLMs failed to produce the desired output. Mistral-7b and Llama3.3-70b, run on 4 NVIDIA A40 GPUs with 46GB VRAM, generated hallucinations, misaligned outputs or correct results only for some polities. In general, these smaller models tend to focus on the last few elements of the input data. Instead the larger models, running on Google cloud, correctly generated the output. We were able to run experiments 2 and 3 only with models larger than 100 billion parameters. The experimental setting for experiment 2 consists of the 5 SDT labels and 93 examples, the same used for evaluating the annota-

tion in the Chronos dataset. We evaluated this experiment with both Fleiss'and Cohen's k on the results generated by Mistral-large-2411 (123b), GPT4 (1.8t) and Llama3.1-405b.

The results, reported in Table 1, reveal that the Inter-Annotator agreement between the largest models (Llama3.1-405b and GPT-4) is comparable to the best human performance, while combinations with smaller LLMs yield poor agreement. This suggests that the number of parameters in the model greatly affects the ability of the LLM to evaluate a complex context and select a label. In order to have comparable results, We also tested the agreement between humans and LLMs, finding that this combination yields the best results. Experiment 3 on Intra-Annotator agreement revealed that Llama3.1-large has great consistency, higher than humans, while GPT-4 proved to have lower consistency.

5 Discussion and Conclusion

The observed relation between model size and annotation agreement in the identification of SDT phases from short texts highlights the profound contextual demands of this task. Our findings suggest that models exceeding 400 billion parameters exhibit significantly improved Inter-Annotator agreement over smaller models, reaching human performance. We suggest that very large LLMs have access to more information than humans, and more work is needed to understand how LLMs use this information for label selection. We plan to address this issue in future work.

In conclusion, our observations reinforce the notion that inherently subjective annotation tasks such as SDT phase recognition, necessitates a vast reservoir of learned knowledge. While larger models offer a great performance, their demand for high computational power raises concerns regarding cost, environmental impact, and accessibility. Unfortunately, smaller, locally deployable models, though potentially more sustainable and democratizing, are more prone to hallucinations and hardly achieve acceptable results in this task. This trade-off between performance and practicality underscores the critical need for further research into efficient knowledge representation and transfer within LLMs, particularly when balancing the need for accuracy with the goal of democratizing access to historical analysis. Under this perspective, universities and companies should collaborate, sharing know-how and computational power.

Acknowledgments. This research was supported by the European Commission, grant 10121294: Bankable by Design, Continuous, and Predictive Climate Adaptation Investments with Co-Benefits - CLIMINVEST.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Celli, F., Basile, V.: History repeats: Historical phase recognition from short texts. Proceedings of the CLIC-it (2024)

2. Celli, F., Mingazov, D.: Knowledge extraction from llms for scalable historical data annotation. Electronics **13**(24), 4990 (2024)
3. Cohen, J.: Statistical power analysis for the behavioral sciences. Academic Press, New York (1977)
4. Fleiss, J.L., Levin, B., Paik, M.C.: The measurement of interrater agreement. Statistical methods for rates and proportions **2**, 212–236 (1981)
5. Ge, Y., Yu, H.T., Lei, C., Liu, X., Jatowt, A., Kim, K.s., Lynden, S., Matono, A.: Implicit knowledge-augmented prompting for commonsense explanation generation. Knowledge and Information Systems pp. 1–36 (2025)
6. Goldstone, J.A.: A theory of political demography. Political demography: How population changes are reshaping international security and National Politics pp. 10–28 (2012)
7. Goldstone, J.A.: Demographic structural theory: 25 years on. Cliodynamics **8**(2) (2017)
8. Korotaev, A.V.: Introduction to social macrodynamics: Secular cycles and millennial trends in Africa. Editorial URSS (2006)
9. Korotayev, A., Zinkina, J.: Egypt's 2011 revolution: A demographic structural analysis. In: Handbook of revolutions in the 21st century: The new waves of revolutions, and the causes and effects of disruptive political change, pp. 651–683. Springer (2022)
10. Orlandi, G., Hoyer, D., Zhao, H., Bennett, J.S., Benam, M., Kohn, K., Turchin, P.: Structural-demographic analysis of the qing dynasty (1644–1912) collapse in china. Plos one **18**(8), e0289748 (2023)
11. Pavlovic, M., Poesio, M.: The effectiveness of llms as annotators: A comparative overview and empirical analysis of direct representation. LREC-COLING 2024 p. 100 (2024)
12. Piao, S., Dallachy, F., Baron, A., Demmen, J., Wattam, S., Durkin, P., McCracken, J., Rayson, P., Alexander, M.: A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. Computer Speech & Language **46**, 113–135 (2017)
13. Piketty, T.: Capital in the twenty-first century. Harvard University Press (2014)
14. Sprugnoli, R., Tonelli, S.: Novel event detection and classification for historical texts. Computational Linguistics **45**(2), 229–265 (2019)
15. Stoev, T., Tonkin, E.L., Yordanova, K., Tourte, G.J.: Tutorial: developing a data annotation protocol. In: Ubicomp/ISWC 2023 (2023)
16. Turchin, P.: Long-term population cycles in human societies. Annals of the New York Academy of Sciences **1162**(1), 1–17 (2009)
17. Turchin, P.: Political instability may be a contributor in the coming decade. Nature **463**(7281), 608–608 (2010)
18. Turchin, P.: A Structural-Demographic Analysis of American History. Beresta Books Chaplin (2016)
19. Turchin, P.: End times: elites, counter-elites, and the path of political disintegration. Penguin (2023)
20. Turchin, P., Korotayev, A.: The 2010 structural-demographic forecast for the 2010–2020 decade: A retrospective assessment. PloS one **15**(8) (2020)
21. Turchin, P., Nefedov, S.A.: Secular cycles. Princeton University Press (2009)
22. 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. Journal of Cognitive Historiography **5**, 115–123 (2020)