Monitoring Historical Cultural Hacking in Large Language Models

Fabio Celli[1,*], Astik Samal[1]

[1]Research and Development, Maggioli SpA, Via Bornaccino 101, Santarcangelo di Romagna, Italy

Abstract

Large Language Models, powerful deep learning systems trained on huge textual datasets to generate and analyze human-like language, are increasingly used as sources of historical and cultural knowledge. An underrated aspect in the field of Trustworthy Artificial Intelligence is that societies dependent on Large Language Models developed in foreign countries are at risk of 'cultural hacking' – the implicit or deliberate misalignment of information potentially leading to new forms of cultural imperialism. This preliminary study investigates cultural hacking by analyzing the historical knowledge encoded in Large Language Models developed in the United States and China. We reveal significant disparities in how these models interpret elite power, with Chinese models emphasizing force and rigid stratification, and US models leaning towards wealth-based power. This research highlights the potential for cultural biases to permeate Large Language Models, emphasizing the need for transparent tools to monitor and mitigate algorithmic cultural imperialism, ensuring trustworthy deployment of Artificial Intelligence in sensitive domains. We suggest that democratic societies could mitigate cultural hacking by developing tools for comparing the output from multiple Large Language Models, fostering critical thinking.

Keywords

Trustworthy AI, Cultural Hacking, Large Language Models, Computational History

1. Introduction

The increasing reliance on Artificial Intelligence (AI) and Large Language Models (LLMs) potentially presents significant cultural vulnerabilities. By centralizing vast amounts of information, LLMs create a single point of failure [1], increasing susceptibility to information manipulation and censorship. Recent research found that Chinese LLMs, when prompted in Simplified Chinese, often adhere to Chinese censorship guidelines, unlike responses to prompts written in Traditional Chinese [2]. Culture and governments can shape the information contained in LLMs in two main ways. On the one hand, culture can indirectly influence the data used for training the models, on the other hand, governments can directly regulate the companies that develop AI. This is especially critical for cultures dependent on LLMs developed in foreign countries, which face the risk of 'cultural hacking' – the implicit or deliberate misalignment of cultural information to serve external interests [3].

Historical knowledge, essential for critical thinking and fact checking [4], is particularly at risk. This raises concerns about the potential for algorithmic cultural imperialism and increases the need for culturally relevant computational methods for transparent and trustworthy AI [5]. In this perspective, a key challenge is the transparency of LLMs, as they often function as "black boxes", making it difficult to understand their internal workings and knowledge. Although it is possible to test the historical knowledge of LLMs about specific events with direct questions [6], it is more difficult to test their high-level interpretation of historical events. Moreover, it is not easy to do research on this topic because historical interpretation is context-demanding, and small models tend to hallucinate when performing that task [7].

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.

[*]Corresponding author.

fabio.celli@maggioli.it (F. Celli); astik.samal@maggioli.it (A. Samal)
https://huggingface.co/facells (F. Celli)
0000-0002-7309-5886 (F. Celli); 0009-0004-5851-6621 (A. Samal)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-12-16

Nevertheless, recent work showed that it is possible to extract and visualize LLMs' historical knowledge by means of Time-Resolved Variable prompting [8] (TRV prompting), a method for compressing LLMs' information in scaled sequences, generating comparable synthetic historical data in tabular format. In particular, Using this method, we aim to explore the differences in historical knowledge of LLMs produced by two different cultures: American and Chinese. In particular we test the following hypotheses:

- H1: *Assuming that a LLM's historical knowledge is shaped by its cultural origins, we would expect that, when given the same prompt, historical data generated by LLMs developed within the same culture will correlate much more than those from LLMs developed in different cultures.*
- H0: *the relationship between historical data points generated by LLMs will not differ significantly, regardless of whether those LLMs originated from different cultures or the same culture.*

The paper is structured as follows: in Section 2 we review related works in the field of Trustworthy AI, in Section 3 we describe the prompt and the experiments, and finally in Section 4 we draw our conclusions.

2. Related Work

The field of Trustworthy AI addresses the critical need for algorithms and systems to be reliable, ethical, and beneficial to societies [9]. This field broadly encompasses two key categories: (i) Fairness, which includes tasks such as bias detection [10] [11] and adversarial robustness [12]; (ii) transparency, focusing mainly on explainability [13] [14] [15].

Going beyond Trustworthy AI, LLMs are already employed for sentiment analysis in the political domain [16] [17]. Crucially, there is a recent trend in investigating both historical and cultural representation in LLMs [18]. Despite the lack of a clear definition of culture, scholars exploit specific cultural proxies measured through specially designed datasets, such as political bias [19] [20]. In general, this emerging field of research on socio-cultural aspects of LLMs has its focus on topics such as value alignment [21], cultural values in LLM adoption [22]; personality synthesis [23] and dilemma resolution in diverse value systems [24] [25], also specifying a cultural identity in the prompt [26]. Current models are known to exhibit different behaviors when the same prompt is queried in different languages [27]. Moreover, there are proofs of strong biases towards Western, Anglocentric, or American cultures, and these biases can lead to reduced utility, unfairness, user alienation or the potential reinforcement of cultural homogeneity [28]. Due to the complex, contextual, and often implicitly understood nature of culture, it is necessary to have a transparent and interdisciplinary approach to understand the culture encoded within LLMs and prevent forms of cultural imperialism [29].

TRV prompting is an information compression technique that is used to represent the evolution of categorical variables over time. It consists in defining chronological sequences of categories, associated to progressive indexes. For instance, in a TRV prompt, the evolution of military technology might use 0.1 for indicating stone/wood, 0.2 for copper, 0.3 for bronze, and so on, building a sequence according to the temporal order of the encoded concepts. The numerical indexes used in TRV prompting can be interpreted both by humans and LLMs, hence can be useful for transparency and model explainability. TRV prompting was shown to be more effective than One-Hot Encoding with Principal Component Analysis in explaining the increase of social complexity from historical synthetic data [30]. Here we propose to apply TRV prompting for generating synthetic historical data from multiple LLMs, that can be compared one against the other. In what follows, we compare the historical data generated by different LLMs.

3. Experiments and Discussion

We generated synthetic historical data with the TRV prompt reported in figure 1. We used the same prompt with six different LLMs, developed in two different countries: the US, and China. In order to represent all model sizes, we used a cluster sampling, with two small, two medium and two large LLMs, one per country. For the small size cluster we selected LLMs with less than 32 billion parameters, for the medium size we selected LLMs around 70 billions parameters, and for the large size we used LLMs above 400 billion parameters. We randomly selected the LLMs within the cluster from a list of models available on the Huggingface playground[1].

Act as an expert historian and generate a table in jsonl format with information about the polities existed in Iran, US, China, Russia and France between 1600 and 2000. the structure of the jsonl must have the following columns:
"time": starting from 1600, proceed with a time-step of 50 years, for example 1600, 1650, 1700 and so on);
"polity": location and name of the polity;
"pop": 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;
"terr": 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;
"soc": social structure type: 0.1=egalitarian/democratic, 0.0=mixed, -0.1=stratified/autocratic, -0.2=totalitarian
"pow": elite power type: 0.0=infrastructure control, 0.1=war/force, 0.2=information/opinion management, 0.3=economy/production/wealth
when values are unknown put null as value.
Below is an example of a row of the desired output:

{ "id": "turkey ottoman empire", "time": "1600", "pop": "0.7", "terr": "0.8", "soc":"-0.1", "pow":"0.2"}

Figure 1: TRV prompt for the generation of synthetic historical data.

The three Chinese models we used are: Deepseek R1, Alibaba's Qwen2 and 01-ai's Yi. Deepseek R1, with 685 billion parameters, has strong reasoning capabilities. It utilizes reinforcement learning and cold-start data to enhance performance and address issues like repetitive outputs. Qwen2 is a 72 billion Instruct model that surpasses many open-source alternatives in various benchmarks, Yi-1.5, with 32 billion parameters, is small but competitive. The three models developed in the US are: Microsoft Phi-4, Perplexity's R1 1776, and Meta's Llama 3.1-70b. Perplexity R1-1776 is a 671 billion parameter reasoning model based on DeepSeek-R1, specifically modified to remove Chinese Communist Party censorship. Llama 3.1 is a family of multilingual text-based models with Grouped-Query Attention (GQA) for improved inference scalability, Phi-4 is a 14 billion parameter model, very small but competitive on common benchmarks. It has been fine-tuned for instruction-following and safety.

model	time	polity	pop	terr	soc	pow
perplexity(671b)	1700	france: kingdom	0.7	0.9	-0.1	0.1
yi1.5(34b)	1600	china: qing dynasty	0.7	0.9	-0.1	0.1
phi-4(14b)	1950	france: french gaullist republic	0.8	0.7	0.1	0.3
llama3.1(70b)	1950	us: united states	0.8	1.0	0.1	0.3
qwen2(72b)	1650	us: colonies	0.1	0.4	0.0	0.0
R1(685b)	1950	iran: pahlavi dynasty	0.7	0.9	-0.1	0.3

Table 1
Examples of synthetic data generated with TRV prompting.

We asked the LLMs to analyze the history of Iran, the US, China, Russia, and France from 1600 to 2000, with 50-year intervals. We focused on four dimensions: population size (pop), territory size (terr), social structure (soc), and elite power (pow). An example of the data generated is reported in Table 1. The data is freely available online for replication studies[2].

By examining the 'polity' column, we were able to detect missing values and hallucinations in the models' output. Our analysis showed that only 1.1% of the data was null, generated by Phi-4 the smallest model in our experiment. Crucially, 8.8% of outputs from Chinese models and 1.8% from US models showed hallucinations. These errors manifested as inaccurate polity-time associations, such as the Qing

[1] https://huggingface.co/playground
[2] https://github.com/facells/fabio-celli-publications/blob/main/docs/chd.zip

hallucinations per country	US models	Chinese models	total
Iran		7 (yi) + 1 (qwen)	8
US		3 (yi)	3
China	1 (llama)	4 (yi) + 1 (qwen)	6
Russia	2 (llama)	6 (yi) + 1 (qwen)	9
France	2 (perplexity)	1 (yi)	3
count	5	24	29
percentage	1.8%	8.8%	10.6%

Table 2
Counts of hallucinations by country and model (denominator = total number of model × country × time step records). Percentages refer to the proportion of hallucinations out of total outputs. The data is aggregated by polity and includes all variables (pop, terr, soc, pow).

dynasty being assigned to the 1600s (see Table 1), and the persistence of a single polity, like Safavid Iran over an unrealistic time span (1600-2000). Details are reported in Table 2.

The population and territory were relatively straightforward, as the information can be retrieved from existing data, and we provided classification instructions as numerical ranges. However, social structure and elite power required more complex reasoning. For these, we provided simplified scales: egalitarian to totalitarian for social structure, and infrastructure control to wealth-based power for the elite power dimension. We also allowed the models to return 'null' values, the model temperature was set on 0.5 and the top-p on 0.7. Results are single shots.

Settings	culture	corr	Settings	culture	corr
deepseek-pow + qwen-pow	cn	0.816**	deepseek-soc + qwen-soc	cn	0.708**
deepseek-pow + yi-pow	cn	0.461**	deepseek-soc + yi-soc	cn	0.785**
qwen-pow + yi-pow	cn	0.613**	qwen-soc + yi-soc	cn	0.672**
deepseek-pow + perplexity-pow	mix	**0.338***	deepseek-soc + perplexity-soc	mix	0.825**
qwen-pow + llama-pow	mix	**0.193****	qwen-soc + llama-soc	mix	0.661**
yi-pow + phi-pow	mix	0.538**	yi-soc + phi-soc	mix	0.809**
perplexity-pow + llama-pow	us	0.463*	perplexity-soc + llama-soc	us	0.743**
perplexity-pow + phi-pow	us	0.565**	perplexity-soc + phi-soc	us	0.827**
phi-pow + llama-pow	us	0.750**	phi-soc + llama-soc	us	0.816**
deepseek-terr + qwen-terr	cn	0.742**	deepseek-pop + qwen-pop	cn	0.522*
deepseek-terr + yi-terr	cn	0.605**	deepseek-pop + yi-pop	cn	0.724**
qwen-terr + yi-terr	cn	0.899**	qwen-pop + yi-pop	cn	0.838**
deepseek-terr + perplexity-terr	mix	**0.266***	deepseek-pop + perplexity-pop	mix	0.781**
qwen-terr + llama-terr	mix	0.839**	qwen-pop + llama-pop	mix	0.813**
yi-terr + phi-terr	mix	0.945*	yi-pop + phi-pop	mix	0.885**
perplexity-terr + llama-terr	us	0.823**	perplexity-pop + llama-pop	us	0.860**
perplexity-terr + phi-terr	us	0.892**	perplexity-pop + phi-pop	us	0.864**
phi-terr + llama-terr	us	0.853*	phi-pop + llama-pop	us	0.873*

Table 3
Pearson correlations for each pair of LLM and dimension. **=*p-value* below 0.01, *=*p-value* below 0.05. Stars mark the statistical significance while bold marks the cases where H1 is supported (mix is lower than cn and us).

Then we tested H1 (whether historical data generated by LLMs developed within the same culture correlates much more closely than those from LLMs developed in different cultures). To do so, we computed the Pearson correlation matrix by excluding null values and comparing the series generated by LLMs from the same (us, cn), acting as control group, and different cultures (mix), the experimental group.

A limitation of this study is that it is possible to control the effect of LLM size only in the experimental group. In order to reject the null hypothesis (H0) we expect that correlations between values generated by LLMs from different cultures are lower than the ones generated by LLMs from the same culture.

LLM	corr to real-terr	corr to real-pop
Deepseek	0.401	0.505
Qwen	0.895	0.961
Yi	0.859	0.861
Perplexity	0.767	0.864
Llama	0.306	0.511
Phi	0.724	0.791

Table 4
Pearson correlations of each LLM to real territory and population data (mapped to TRVs).

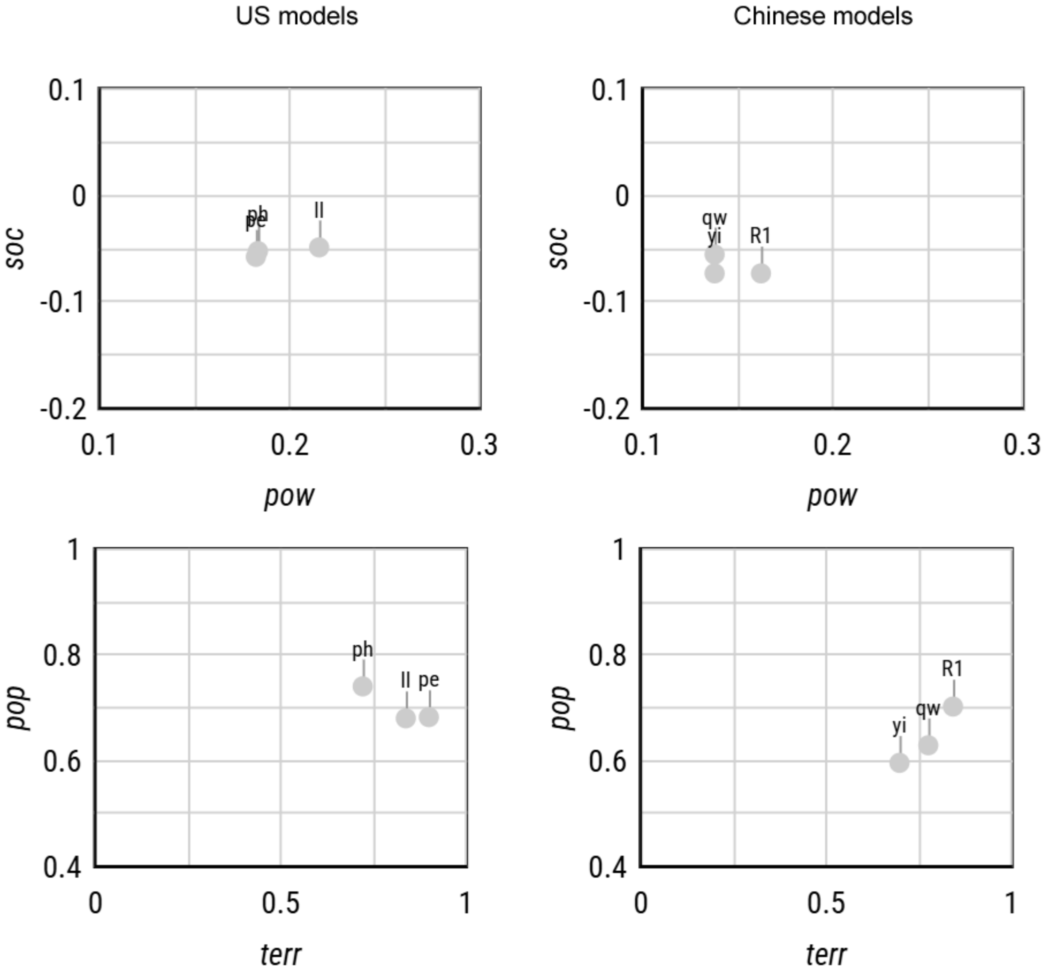

Figure 2: Visualization of historical predictions of US and Chinese LLMs averaged by model. Legenda: ph=Phi4, ll=Llama3.1, pe=Perplexity, R1=Deepseek, qw=Qwen2, yi=Yi; social structure type (soc) 0.1=democratic, 0=mixed, -0.1=autocratic, -0.2=totalitarian; elite power type (pow): 0.1=force, 0.2=opinion management, 0.3=wealth; polity population (pop): 0.5=between 50 and 100 thousand people, 0.7=between 1 and 100 million people, 0.9=between 500 million and 1 billion people; polity territory (terr): 0.1=less than 100 km2, 0.5=between 10 and 50 thousand km2, 1.0=more than 10 million km2.

Results, reported in table 3, reveal that the correlations between values generated by LLMs from different cultures (mix) are lower than those from the same culture (cn, us) only in the case of Elite power structure (pow) and territory size (terr), and only with the largest models. This allows us to reject the null hypothesis only in these two cases. Crucially, for 'soc' (social structure) and 'pop' (population), the 'mix' correlations are not consistently lower and, in some cases, are quite high (e.g., deepseek-soc + phi-soc: 0.968, deepseek-pop + perplexity-pop: 0.781).

As expected we observe the lowest correlations in a subjective dimension such as elite power

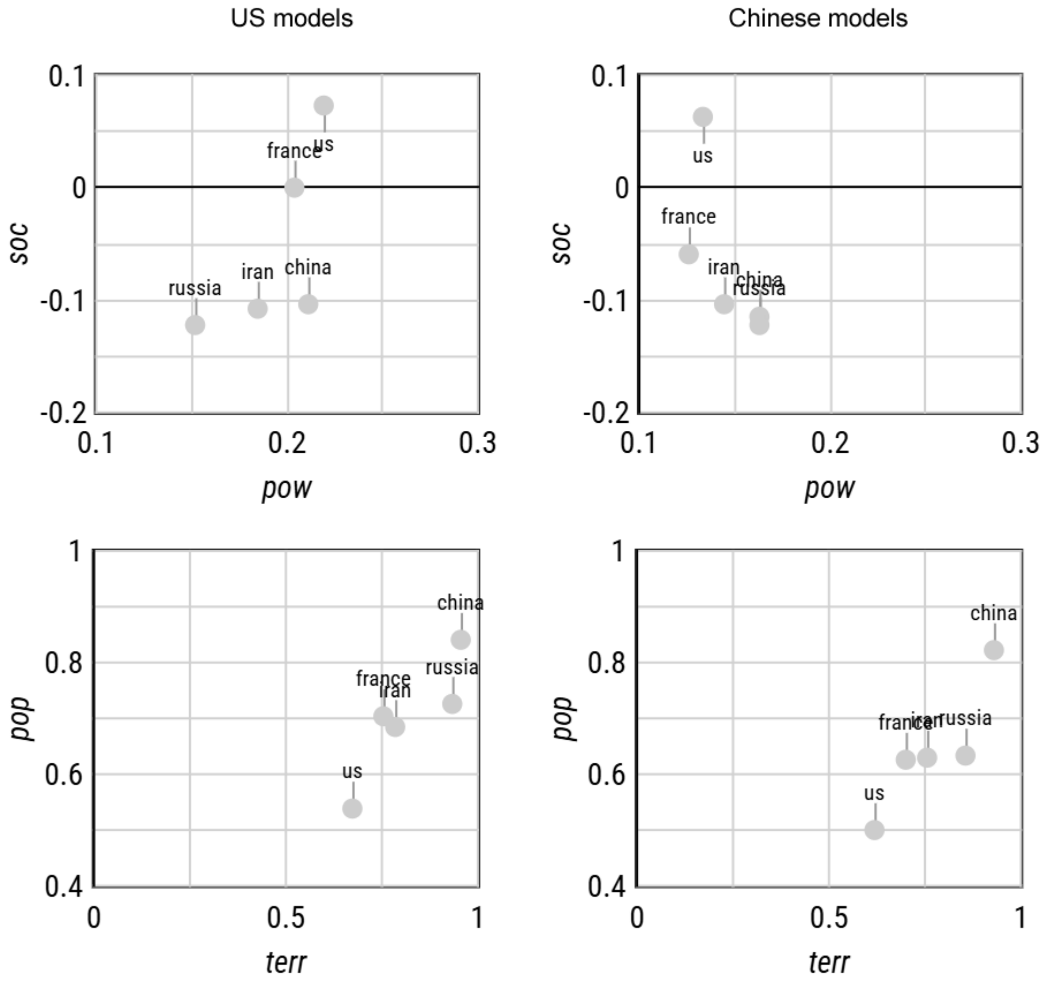

Figure 3: Visualization of historical predictions of US and Chinese LLMs averaged by country. Legenda: social structure type (soc) 0.1=democratic, 0=mixed, -0.1=autocratic, -0.2=totalitarian; elite power type (pow): 0.1=force, 0.2=opinion management, 0.3=wealth; polity population (pop): 0.5=between 50 and 100 thousand people, 0.7=between 1 and 100 million people, 0.9=between 500 million and 1 billion people; polity territory (terr): 0.1=less than 100 km2, 0.5=between 10 and 50 thousand km2, 1.0=more than 10 million km2.

structure, while it is strange to observe a poor correlation in a measurable dimension such as territory. A comparison to real data (mapped to TRVs) reveals that Llama and Deepseek have the poorest correlations to both territory and population scores, as reported in Table 4. This suggests that cultural hacking might affect both abstract and measurable concepts, also in the largest models. We then visualized and compared the data with scatterplots on a dashboard. The goal is to have visual cues of how much the LLMs' country of origin affected the historical information they produced.

As shown in Figure 2, the Chinese models consistently assigned lower values to elite power and social structure compared to the US models. This suggests the Chinese models perceive societies as more rigidly stratified and autocratic, and see elite power as stemming more from force and manipulation of public opinion. Conversely, the US models, particularly Llama, were less likely to interpret societies as highly stratified and leaned towards attributing power to wealth. This difference highlights how LLMs can reflect the cultural context in which they are developed.

When analyzing the scatterplots of population and territory size, we observed a general agreement among the US models. However, the Chinese models showed more variability, and we suggest this is due to the high number of hallucinations in the smaller models, Yi1.5 and Qwen2. Notably, the R1 model's results aligned with those of the US models.

We also visualized the predictions of US and Chinese models averaged by country. This revealed distinct interpretations of elite power that explain the misalignments in correlations. The models developed in the US generally portrayed elites as driven by opinion management and wealth, whereas Chinese models focused on force, as illustrated in Figure 3. Chinese models also exhibited a slight tendency to underestimate population, possibly due to higher hallucination rate of the Chinese models. Considering that most errors are on Russia and Iran, the Chinese and US models differ mostly in their historical interpretation of France.

4. Conclusion

By comparing the outputs of LLMs produced in US and China, this study highlights the risk of cultural hacking, particularly in a sensitive domain like historical interpretation, which is the basis for shared values in societies. The intent of this work is comparative, not normative. We acknowledge that there are some uncontrollable technical factors (e.g. model size in the control group, training data) and regulatory constraints that may overlap with cultural influences, hence we report our findings with this caution in mind. In conclusion, this study shows the feasibility of employing TRV prompting to systematically analyze and compare the historical knowledge encoded within LLMs from different cultural and geopolitical origins. We were able to demonstrate that, given the same prompt, historical data generated by US and Chinese LLMs revealed significant disparities in their interpretations of historical elite power structures. Specifically, Chinese models tended to emphasize force and rigid social stratification, while US models leaned towards wealth-based power and less stratified societies. These differences underscore the areas where cultural hacking is more prone to permeate LLMs, reflecting the contexts in which they are developed. Furthermore, we observed that smaller Chinese models exhibited a higher propensity for hallucinations, impacting their consistency in population and territory estimations.

Beyond the demonstrated disparities and potential for hallucinations, the diffusion of distant writing, the practice of producing literary texts using LLM [31], increases the potential for cultural hacking. As individuals and organizations increasingly rely on LLMs developed in foreign cultures for generating text, there is a significant risk that subtle, or even overt, cultural biases embedded within these models could influence narratives, shape perceptions, and potentially erode the shared understanding of historical events or societal values. This could lead to a gradual, unconscious shift in perspectives, as the interpretations favored by the foreign LLMs become normalized through their widespread use in content creation, education, and potentially even in policy discussions. This necessitates not only critical evaluation of LLM outputs but also the development of tools that allow users to identify and mitigate these potentially pervasive, yet often invisible, influences. While our monitoring was primarily based on correlations and scatterplots to identify differences, in the future it is possible to explore other quantitative methods, such as cluster analysis and distance metrics, to rigorously quantify and compute cultural divergences in LLM knowledge, and also develop monitoring systems to track and visualize cultural changes in LLMs over time. Addressing the "black box" nature of LLMs and developing tools to monitor cultural hacking is crucial for mitigating the risks of algorithmic cultural imperialism and ensuring the responsible deployment models produced abroad. As LLMs become increasingly integrated into information ecosystems, understanding and mitigating their potential cultural biases is essential for safeguarding cultural security and promoting trustworthy AI. In particular, we suggest that democratic societies could mitigate cultural hacking by using TRV prompting and visualization techniques to compare the output from multiple LLMs, fostering critical thinking. We believe that this research contributes to the broader agenda of enhancing human trust in AI systems by exposing and interrogating the implicit cultural biases within LLMs. The methodological framework we propose offers an accessible approach to knowledge extraction, that can be exploited also by non-experts. This goes in the direction of a multidisciplinary approach to the problem. By surfacing the often opaque historical and cultural assumptions embedded in these models, this research supports the development of more human-centric AI systems, where cultural context and societal values are integral to model evaluation and design.

Acknowledgments

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

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini 2.0 in order to do grammar and spelling check. The authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Ethics Statement

This research involved the analysis of publicly available Large Language Models (LLMs) developed in the United States and China. Although no direct interaction with human participants was conducted, we acknowledge the sensitive nature of comparing AI models developed in different geopolitical contexts and have strived for objectivity in our analysis and interpretation of the results. We are committed to transparent reporting of our methodology, limitations and findings to facilitate critical evaluation and further research in this important area.

References

[1] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[2] M. Ahmed, J. Knockel, The impact of online censorship on llms, Free and Open Communications on the Internet (2024).

[3] C. O'neil, Weapons of math destruction: How big data increases inequality and threatens democracy, Crown, 2017.

[4] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, A. Flammini, Computational fact checking from knowledge networks, PloS one 10 (2015) e0128193.

[5] R. J. Deibert, Reset: Reclaiming the internet for civil society, House of Anansi, 2020.

[6] J. Hauser, D. Kondor, J. Reddish, M. Benam, E. Cioni, F. Villa, J. S. Bennett, D. Hoyer, P. Francois, P. Turchin, et al., Large language models' expert-level global history knowledge benchmark (hist-llm), in: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024.

[7] F. Celli, V. Basile, Large language models rival human performance in historical labeling, in: Proceedings of ARDUOUS 2025, Co-located with ECAI 2025, Bologna, Italy, 2025, p. 5.

[8] F. Celli, D. Mingazov, Knowledge extraction from llms for scalable historical data annotation, Electronics 13 (2024) 4990.

[9] D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy artificial intelligence: a review, ACM computing surveys (CSUR) 55 (2022) 1–38.

[10] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM computing surveys (CSUR) 54 (2021) 1–35.

[11] P. Salmani, P. R. Lewis, Transfer learning can introduce bias, in: ECAI 2024, IOS Press, 2024, pp. 2426–2433.

[12] E. Lee, Y. Lee, T. Lee, Adversarial attack-based robustness evaluation for trustworthy ai., Computer Systems Science & Engineering 47 (2023).

[13] G. Zaza, G. Casalino, G. Castellano, The first workshop on explainable artificial intelligence for the medical domain-explimed@ ecai2024 (2024).

[14] J. M. Alonso-Moral, Z. Anthis, R. Berlanga, A. Catalá, P. Cimiano, P. Flach, E. Hüllermeier, T. Miller, D. Mindlin, O. Mitruț, et al., Preface of the proceedings of the workshop on multimodal, affective and interactive explainable artificial intelligence (mai-xai), collocated with the european conference on artificial intelligence (ecai) (2024).

[15] M. Gniewkowski, P. Walkowiak, P. Syga, M. Klonowski, T. Walkowiak, Do not trust me: Explainability against text classification, in: ECAI 2023, IOS Press, 2023, pp. 875–882.

[16] L. S. Zaabar, A. A. Yacob, M. R. Mohd Isa, M. Wook, N. A. Abdullah, S. Ramli, N. A. M. Razali, Sentiment and emotion analysis with large language models for political security prediction framework., International Journal of Advanced Computer Science & Applications 16 (2025).

[17] J. T. Ornstein, E. N. Blasingame, J. S. Truscott, How to train your stochastic parrot: Large language models for political texts, Political Science Research and Methods 13 (2025) 264–281.

[18] F. De Ninno, M. Lacriola, Mussolini and chatgpt. examining the risks of ai writing historical narratives on fascism, Journal of Modern Italian Studies (2025) 1–23.

[19] S. Feng, C. Y. Park, Y. Liu, Y. Tsvetkov, From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 11737–11762. doi:10.18653/v1/2023.acl-long.656.

[20] L. Rettenberger, M. Reischl, M. Schutera, Assessing political bias in large language models, Journal of Computational Social Science 8 (2025) 1–17.

[21] P. Vongpradit, A. Imsombut, S. Kongyoung, C. Damrongrat, S. Phaholphinyo, T. Tanawong, Safecultural: A dataset for evaluating safety and cultural sensitivity in large language models, in: 2024 8th International Conference on Information Technology (InCIT), IEEE, 2024, pp. 740–745.

[22] S. Lambiase, G. Catolino, F. Palomba, F. Ferrucci, D. Russo, Investigating the role of cultural values in adopting large language models for software engineering, ACM Transactions on Software Engineering and Methodology (2024).

[23] F. Celli, A. Kartelj, M. Đorđević, D. Suhartono, V. Filipović, V. Milutinović, G. Spathoulas, A. Vinciarelli, M. Kosinski, B. Lepri, Twenty years of personality computing: Threats, challenges and future directions, arXiv preprint arXiv:2503.02082 (2025).

[24] M. Mozikov, N. Severin, V. Bodishtianu, M. Glushanina, I. Nasonov, D. Orekhov, P. Vladislav, I. Makovetskiy, M. Baklashkin, V. Lavrentyev, et al., Eai: Emotional decision-making of llms in strategic games and ethical dilemmas, Advances in Neural Information Processing Systems 37 (2024) 53969–54002.

[25] A. Jha, P. Mann, A. Tiwari, K. Kadian, A. Sharma, Decoding ethics: Proficiency of llms in addressing moral dilemmas, in: The International Conference on Recent Innovations in Computing, Springer, 2023, pp. 593–605.

[26] Y. Tao, O. Viberg, R. S. Baker, R. F. Kizilcec, Cultural bias and cultural alignment of large language models, PNAS nexus 3 (2024) pgae346.

[27] J. G. Lu, L. L. Song, L. D. Zhang, Cultural tendencies in generative ai, Nature Human Behaviour (2025) 1–10.

[28] H. R. Kirk, A. Whitefield, P. Rottger, A. M. Bean, K. Margatina, R. Mosquera-Gomez, J. Ciro, M. Bartolo, A. Williams, H. He, et al., The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models, Advances in Neural Information Processing Systems 37 (2024) 105236–105344.

[29] M. F. Adilazuarda, S. Mukherjee, P. Lavania, S. Singh, A. F. Aji, J. O'Neill, A. Modi, M. Choudhury, Towards measuring and modeling" culture" in llms: A survey, arXiv preprint arXiv:2403.15412 (2024).

[30] F. Celli, How to compress categorical variables to visualize historical dynamics, in: Proceedings of the 21st Conference on Information and Research science Connecting to Digital and Library science, 2025, pp. 1–13.

[31] L. Floridi, Distant writing: Literary production in the age of artificial intelligence, Minds and Machines 35 (2025) 1–26.