MC2024
PEOPLE & BUILDINGS
Proceedings of 13th Masters Conference:
London, 16th September 2024

Network for Comfort &
Energy Use in Buildings
http://www.nceub.org.uk

How much do Large Language Models know about panel paintings preservation?

Fengjun Li[1], Dr. Daniela Reggio[2] and Dr. Josep Grau-Bove[3]

[1] MSc Built Environment: Sustainable Heritage (Data Science), University of London, UK, fengjun.li.23@ucl.ac.uk
[2] Institute for Sustainable Heritage, University College London, London, UK, daniela.reggio@ucl.ac.uk
[3] Institute for Sustainable Heritage, University College London, London, UK, josep.grau.bove@ucl.ac.uk

Abstract: This study evaluates the capability of large language models (LLMs) in understanding the preservation of paintings on panel by comparison with the predictions obtained through the digital platform HERIe. The latter is specialized tool for heritage object risk assessment. Four large language models (LLMs) - ChatGPT 3.5, ChatGPT 4, Claude, and Gemini - were tested asking what are the levels of strain experienced by panel paintings under different conditions. The models were also tested on their ability to rank different environments conditions in order of suitability for storing panel paintings and were examined whether the languages of prompts affected results. The study concludes that while LLMs demonstrate a general understanding of wood panel preservation principles, they lack the specialized calculation abilities of purpose-built tools like HERIe for precise risk assessment in cultural heritage preservation.

Keywords: panel paintings, preservation, microclimate, LLMs

1. Introduction

The conservation of wooden panel paintings is an important issue worldwide, although traditional methods have achieved certain results, they still face many challenges, such as technical limitations and high costs. With the emergence of large language models (LLMs), it has brought revolutionary impacts to the field of cultural heritage, which demonstrated potential value in damage prediction. This study aims to evaluate the effectiveness and accuracy of LLMs in practical applications by comparing the performance of LLMs and the computational assessment tool HERIe in predicting the strain of panel paintings. Through data analysis and simulated case studies, this article will explore the practical application prospects and potential limitations of LLMs in the field of paintings preservation.

2. Literature Review

There is a growing body of research examining the knowledge level of LLMs in various fields, such as archaeology (Agapiou and Lysandrou, 2023), obstetrics (Sarraju et al., 2023), programming (Surameery and Shakor, 2023), preservation of traditional architecture (Zhang, 2024) and water management (Emenike et al., 2023). The rapid rise of Artificial Intelligence (AI) has brought more possibilities for development of various fields including the cultural heritage sector. Spennemann (2023) explored how genAI language models present cultural heritage and explored ChatGPT's capabilities of curation. By comparing with experimental data from digitized sites, Yilmazer & Karakose (2023) concluded that ChatGPT can quickly provide valuable information in the field of cultural heritage. However, compared to disciplines such as medicine or education, there are few papers exploring the application of large language models in the heritage field. Currently, these papers mainly explore the blogging domain in cultural heritage environments, museum management, and value interpretation. There is little discussion on the capabilities of LLMs in paintings preservation, including the preservation of panel paintings.

3. Methodology

This study drew on four publicly accessible large language models (ChatGPT 3.5, ChatGPT 4, Gemini, and Claude 3.5 Sonnet) to engage in a 'conversation' with each as to how much do LLMs know about paintings conservation and preservation. Based on the papers reviewed (Agapiou and Lysandrou., 2023) and the purpose of this paper, I conducted a series of queries on LLMs related to panel paintings conservation and preservation.

4. Results and Discussion

Question one: *"In our conversation, please think as an expert on panel paintings and answer my questions. The csv file I provide you records the temperature and relative humidity values of a museum room for a whole year. In this environment, a 10 mm thick unrestrained panel painting covered with 0.4 mm gesso layer is stored. Please predict the strain of this panel painting at each time point under such environmental conditions."*

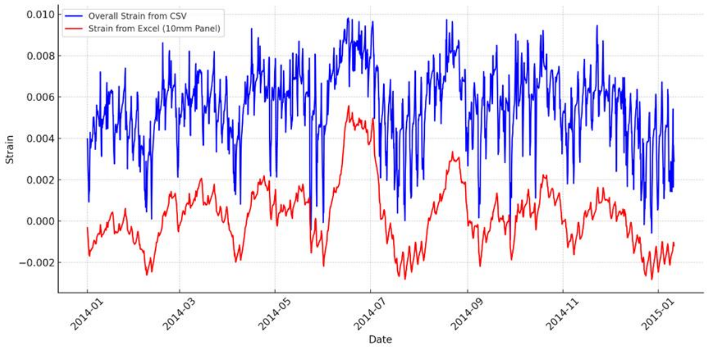

Figure 1. comparison between gpt-4 and HERIe(strain)

The blue one represents the prediction of gpt-4 and the red line represents the calculation of HERIe. The chart shows that at the same time point, gpt-4 believes that the strain experienced by panel paintings is always greater than the result calculated by HERIe. HERIe uses the Mecklenburg's approach when calculating the strain to changes in temperature and RH of the objects, which is the worst-case scenario. This shows that gpt-4 believes that the strain that the object will experience is worse than the worst-case scenario. Therefore, the calculation results from GPT-4 have no reference value. It is exciting that the two lines have the same pattern, proving that part of gpt-4's judgment overlaps with HERIe's design. However, in half of the cases, gpt-4 cannot judge whether the strain is positive or negative. Therefore, it can be determined that there is a large gap between gpt-4 and HERIe in the judgment of specific strain values.

Question 2: *"You are a conservator at a museum specializing in paintings on panel. The data I am providing you now are the changes in the indoor temperature and relative humidity of a museum within a year. There are three types of wood panels in this museum. The first type is a 10 mm wooden panel covered with a gesso layer, and the second type is a 40 mm wood panel covered with a gesso layer. The surface of these two types of wood panels is covered with a 0.4 mm gesso layer. This stiff 0.4 mm thick gesso layer laid on a tangentially cut oak panel open to a water vapor flow through both faces of the panel. The third type is a fully restrained 20 mm thick tangentially cut oak panel open to a water vapor flow through both faces of the panel. Based on the temperature and humidity data I provided to you, please* *predict whether these three types of wood panels will experience strain under such environmental conditions and rank the four situations according to the severity of the strain experienced"*

Since other models do not support attaching data files, the data sets were described instead to allow the LLMs to evaluate them. Data description based on the statistical analysis of four datasets (Table 1). This method was also used to test GPT-4. Such a comparison can determine whether describing the data set can serve as an alternative to sending the actual data set.

Table 1. Statistical analysis of four scenarios

Scenario	Temperature Mean (C)	Temperature standard deviation	Humidity Mean (%)	Humidity standard deviation
1	19.95	1.15	49.89	2.90
2	19.95	1.15	52.32	12.64
3	17.50	8.85	49.89	2.90
4	13.20	5.96	68.47	5.98

Severity ranking for all Panel Types is Scenario 1, Scenario 3, Scenario 4, most severe: Scenario 2. According to the results of HERIe, severity ranking is 1, 3, 2, most severe: Scenario 4. The principle of HERIe to calculate risk index is that when the absolute maximum strain experienced by the object does not exceed 0.2%, the risk index is zero. When the absolute strain it experiences exceeds 0.4%, the risk index is 1. In Scenario 4, both unrestrained wooden panels have a higher risk of experiencing mechanical damage. Although the risk indexes of mechanical damage of the restrained wooden panel are 0 in all four cases, it also experiences strain. Table 2 shows the ranking results of five iterations of four different large language models:

Table 2. Ranking iterations of LLMs

iterations	gpt4	gpt3.5	Claude	Gemini
1	1342	1324	1342	1324
2	1342	4132	1342	1324
3	1342	1342	1342	1423
4	1342	1324	1342	1342
5	1342	1324	1342	1324

It can be found that only gpt-3.5 and Gemini has given the same answer as HERIe during the iteration process. The performance of these two models is very unstable; although gpt-4 and Claude have never given the same ranking as HERIe, they have always been stable. Obviously, the stability of gpt-3.5 and Gemini is not as good as gpt-4 and Claude. They identified Scenario 2 as the least suitable for panel paintings preservation, citing the reason that the relative humidity in Scenario 2 fluctuates too widely, reaching up to 100%. Although this peak humidity doesn't last long, such drastic, sudden changes in humidity can cause the most severe strain and irreversible mechanical damage. However, HERIe considers Scenario 4 the least suitable for the preservation of panel paintings, as it emphasizes the impact of a long-time changes in relative humidity.

Scenario 1 was thought as the most suitable environment for preserving panel paintings. In the conditions of Scenario 1 and Scenario 3, the risk of mechanical damage to the three types of panel paintings is zero. HERIe provides ratings based on whether the microclimate data meet ASHRAE specifications, rating Scenario 1 as Class B and Scenario 3 as Class C. The difference between Scenario 1 and Scenario 3 is that the temperature fluctuations in Scenario 3 are more pronounced and have a wider range. GPT-4 and Claude detected this, indicating that LLMs can assess the impacts of both temperature and relative humidity. Compared to Scenario 3, Scenario 2 has significant fluctuations in relative humidity but stable temperature, and the LLMs accurately detected that the impact of relative humidity far outweighs that of temperature. At last, compared with the attached data set, the ranking given by gpt-4 based on simple data description is the same, which demonstrates that LLMs can extract information effectively through provided data descriptions.

Question 3: *‟Please rank the fragility of these three types of panel paintings “*

The first is that the all the models think that the thinner the panel is, the more susceptible it is under rapid changes in moisture and temperature due to its lower mass and inertia. The second is that they think being fully restrained implies that this panel cannot move freely in response to dimensional changes due to moisture and temperature fluctuations. This restriction increases the internal stress significantly, potentially leading to cracking or warping, especially given that it is exposed to moisture flow and is cut tangentially. Therefore, under these four environmental conditions, each time all the models considered the second most vulnerable object is the third type of wood panel – Restrained wooden panel of 20 mm thick. However, HERIe's calculation results show that the restrained wooden panel is the strongest type of panel, because in these four environments, HERIe shows that the restrained wooden panel will experience a mechanical damage risk index of 0. The four LLMs give the rankings of 10 mm Wooden Panel with Gesso, 20 mm Fully Restrained Panel and 40 mm Wooden Panel with Gesso. LLMs are lacking in judging the characteristics of objects, when it comes to structural characteristics. As a result, LLMs have a better understanding of unrestrained wooden panels, leading to more accurate judgments.

Question 4: Starting with the relative humidity of 50%, and increasing in 5% intervals, I tested how LLMs assess the strain level at each stage as the relative humidity changes from 50% to 80% and at a constant temperature of 20°C. Table 3 shows the evaluation of the four LLMs and HERIe on the level of strain experienced by 10 mm Wooden panel under different relative humidity conditions and whether it will cause mechanical damage.

Table 3. strain level and mechanical damage (10 mm)

	Gpt-4		Gpt-3.5		Claude		Gemini		HERIe	
Relative humidity	Strain level	Mechanical damage	Strain level	Mechanical damage	Strain level	Mechanical damage	Strain level	Mechanical damage	Strain level	Mechanical damage
50%	None	No	Slight	No	Slight	No	moderate	Yes	Slight	No
55%	Moderate	Yes	Slight	No	Slight	No	moderate	Yes	Slight	No
60%	Moderate to Severe	Yes	Slight	No	Moderate	No	moderate	Yes	Moderate	No
65%	Severe	Yes	Slight	No	Moderate	No	moderate	Yes	Moderate	Yes

70%	Severe	Yes	Slight	No	Severe	Yes	moderate	Yes	Severe	Yes
75%	Severe	Yes	Slight	No	Severe	Yes	moderate	Yes	Severe	Yes
80%	Severe	Yes	Slight	No	Severe	Yes	moderate	Yes	Severe	Yes

Taking HERIe's judgment as the standard, the same as HERIe's judgment is recorded as accurate. Then the accuracy rate and average accuracy rate of the four models for three different wooden panels are shown in the following table:

Table 4. accuracy rate of four models

	gpt4	gpt3.5	Claude	Gemini	Average accuracy rate
10mm Wooden panel covered with a gesso layer	0.64	0.64	0.93	0.36	0.64
40mm Wooden panel covered with a gesso layer	0.57	0.43	0.71	0.36	0.52
Restrained wooden panel (20 mm thick)	0.07	0.93	0.29	0	0.32
Average accuracy rate	0.43	0.67	0.64	0.24	0.49

The table shows that GPT-3.5 has the highest average accuracy rate among the four models, followed by Claude, with Gemini having the lowest. Although GPT-3.5's accuracy rate is higher than Claude's, Tables 6 and 7 indicate that it consistently sets the response to "No" regarding whether mechanical damage occurs to both types of wooden panels at all relative humidity levels. However, the fact is that when the relative humidity reaches 65%, the unrestrained wooden panel does experience mechanical damage. Since this type of damage is irreversible, GPT-3.5's judgment is considered overly optimistic. If applied to the actual protection of wooden panels, it could lead to significant losses. Therefore, the judgment of Claude is more trustworthy.

Table 4 also presents the average accuracy rate of the four models of assessing a specific wooden panel. It can be observed that they have the least understanding of the restrained wooden panel (20 mm thick), resulting in the lowest average accuracy rate, which is even less than 50%. This is consistent with the previous conclusions. It can be concluded that the more complex the structure, the lower the accuracy of the model. All four models consistently agree that the restrained wooden panels are susceptible to damage. They believe that fully restrained panels cannot expand due to the physical constraints that lock them in place. When the wood attempts to expand against these fixed boundaries, the restriction generates internal stress. This accumulated stress may exceed the mechanical strength of the material, leading to structural failures such as cracking or splitting. However, HERIe suggests that mechanical damage will not occur in fully restrained wooden panels if the absolute maximum strain experienced throughout the strain history does not exceed 0.5%. For unrestrained panels, however, this threshold is 0.2%, indicating that restrained wooden panels are more adaptable than unrestrained ones. They can remain stable across a broader range of relative humidity because the material can freely expand and then be "compressed" back to its original restrained length, thereby reducing the risk of mechanical damage. I can conclude that LLMs consider restraint to be a weakness, but in reality, restraint serves as protection.

5. Conclusion

Among the LLMs used in this study, GPT-4 and Claude demonstrated higher stability. While GPT-3.5 and Gemini occasionally provided surprising answers, they were less stable compared to GPT-4 and Claude. When interacting with LLMs, the quality of the prompts is crucial. Although this study found that the results generated from both Chinese and English prompts were identical, this does not necessarily indicate their broad generalization capability in terms of language. The issue raised in Query B indicates that the provision of a dataset does not significantly affect the ranking results; a comprehensive description of the dataset is sufficient. This suggests that, compared to HERIe, the advantage of LLMs is that they can provide accurate judgments without the need for specific data, saving time. Therefore, they can play an important role in preliminary decision-making.

The LLMs have a relatively comprehensive understanding of panel painting conservation. Besides the points mentioned above, they also recognize that relative humidity has a much greater impact on objects than temperature, which aligns with HERIe's calculation principles. However, the LLMs' understanding of the characteristics of different panel structures is inadequate, especially when it comes to restrained wooden panels, where their understanding is much less comprehensive compared to their understanding of unrestrained wooden panels.

When calculating the specific strain values that wooden panels might experience, the results from LLMs differ significantly from those of HERIe. The computational approach of LLMs tends to focus more on point-to-point changes when assessing the strain and mechanical damage experienced by wooden panels, whereas HERIe considers the cumulative effects over time. This indicates that the LLMs' assessments are not sufficiently comprehensive.

These limitations show that although LLMs provide certain help and convenience in the field of woodcut painting conservation, they cannot completely replace traditional conservation methods and professional judgment. LLMs can assist researchers in making preliminary judgments, and museum staff who are not good at wooden panels conservation can also use LLMs to aid in decision-making. LLMs can play a significant role in assessing preservation conditions.

6. References

Agapiou, A. and Lysandrou, V., 2023. Interacting with the artificial intelligence (AI) language model ChatGPT: A synopsis of earth observation and remote sensing in archaeology. Heritage, 6(5), pp.4072-4085.

Emenike, T.I., Anidiobi, S.U. and Adeniyi, A.G., Revolutionizing Water Treatment, Conservation, and Management: Harnessing the Power of AI-Driven ChatGPT Solutions Abel Egbemhenghe1, 2, Toluwalase Ojeyemi3, 4, Kingsley O. Iwuozor5, 6, Ebuka Chizitere.

Sarraju, A., Bruemmer, D., Van Iterson, E., Cho, L., Rodriguez, F. and Laffin, L., 2023. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. Jama, 329(10), pp.842-844. 2000. Balham: CIBSE

Spennemann, D.H., 2023. Exhibiting the heritage of Covid-19—a conversation with ChatGPT. Heritage, 6(8), pp.5732-5749.

Spennemann, D.H., 2023. ChatGPT and the generation of digitally born “knowledge”: How does a generative AI language model interpret cultural heritage values?. Knowledge, 3(3), pp.480-512.

Surameery, N.M.S. and Shakor, M.Y., 2023. Use chat gpt to solve programming bugs. International Journal of Information Technology and Computer Engineering, (31), pp.17-22.

Yilmazer, M. and Karakose, M., 2023, June. Chat with ChatGPT: Access to cultural heritage with ChatGPT. In 2023 International Conference on Sustaining Heritage: Innovative and Digital Approaches (ICSH) (pp. 131-135). IEEE.

Zhang, J., Xiang, R., Kuang, Z., Wang, B. and Li, Y., 2024. ArchGPT: harnessing large language models for supporting renovation and conservation of traditional architectural heritage. Heritage Science, 12(1), p.220.