LLM4Art: Using Large Language Models for Art Interpretation

Zhaohui Li[1]*, Naiquan Li[1]*, Dancheng Liu[1], Yuting Hu[1], Chenhui Xu[1], Shenghao Chao[2], Jinjun Xiong[1]
[1]Univerisity at Buffalo, {zli253; naiquanl; dliu37; yhu54; cxu26; jinjun}@buffalo.edu
[2] Savannah College of Art and Design, shenghaochao20@gmail.com

ABSTRACT: This study introduces LLM4Art, a framework that uses large language models (LLMs) to interpret visual artworks and support public art education. With recent advances in LLMs, we explore how LLMs can analyze visual art to enhance public art understanding and educational engagement. We propose a system that allows users to upload an artwork and receive interpretations along three dimensions: Visual Analysis, Historical Context, and Cultural Significance. We co-designed prompts and rubrics with art scholars to capture interpretive depth beyond surface-level description. To evaluate model performance, we tested five mainstream LLMs from the GPT and Gemini families on a curated dataset. Model outputs were rated by GPT-5, used as an automated judge due to its superior performance. Results show that Gemini models have strong interpretive accuracy, particularly in analyzing well-known artworks. Future work will focus on expanding the dataset, incorporating comprehensive human evaluations, and developing more reliable and educationally aligned LLMs for art understanding.

Keywords: Multimodal LLMs, Art Education, Educational Technology, Artificial Intelligence

1 Introduction

LLMs have shown promising visual reasoning capabilities. Recent studies have explored their performance in the context of visual art interpretation, examining how LLMs analyze paintings (Arita et al., 2025; Khadangi et al., 2025; Bin et al., 2024; Yuan et al., 2023). Researchers also leveraged LLMs to develop art-related applications, such as intelligent museum guides (Trichopoulos et al., 2023). However, there remains a gap in applying LLMs within art education, particularly in evaluating their ability to interpret diverse types of artworks and support learning in instructional settings. To address these challenges, we first co-designed a rubric and prompting strategy that enabled LLMs to evaluate artworks across three educational dimensions: *Visual analysis, Historical context, and Cultural significance* (School of Art and Design, n.d.; Piper, 2016). We then curated a structured dataset of 50 artworks across 10 categories. The dataset is used to assess the interpretation capabilities of LLMs from the GPT and Gemini families in our experiments. To ensure a fair comparison, a benchmark model, GPT-5, serves as a judge, rating the generated analyses. To verify GPT-5's reliability as a judge, we conducted a human–AI agreement test. Finally, we implemented this evaluation pipeline as a web-based application, LLM4Art, for educational art analysis use. In summary, our contributions include: (1) co-designed an LLM Art Interpretation framework with an expert-informed rubric and prompts; (2) curated a multi-category art dataset to evaluate mainstream LLMs' art understanding; (3) conducted comparative experiments to benchmark interpretive performance across visual, historical, and cultural dimensions.

2 Experiment and Results

Data and Model: With guidance from art scholars, we built a small but diverse dataset drawn from publicly available sources such as WikiArt and Smarthistory. The collection (N = 50) has ten art

*These authors contributed equally and share first authorship.

categories, including European Renaissance painting, Ancient Egyptian tomb painting, and Chinese drawing. Each category has five representative works selected to capture a range of periods, styles, and cultural traditions. Each artwork was analyzed by five leading multimodal language models, including GPT-4o, GPT-4o-mini, GPT-o4-mini, Gemini 2.0 Flash, and 2.5 Flash. We use the same set of co-designed prompts to ensure consistency. **Co-design Evaluation Rubric:** In collaboration with art scholars (School of Art and Design, n.d.; Piper, 2016), we identified three analytic dimensions for art interpretation: 1) *Visual Analysis* assesses the accuracy and richness of descriptions related to genre, color, line, shape, and artifact type; 2) *Historical Context* measures the factual correctness and relevance of references to the artwork's period, artist, and stylistic influences; 3) *Cultural Significance* evaluates the depth of insight into meaning, symbolism, and broader social or philosophical impact. Each dimension was rated on a five-point scale (From 1 = incorrect or irrelevant to 5 = comprehensive and insightful).

Evaluation Procedure: We first ran five LLMs on the artworks in our dataset to generate evaluation results. Those model outputs were exported in JSON format and evaluated by GPT-5 using the co-designed rubric to score each LLM's written analysis. GPT-5 was chosen for its strong reasoning ability and limited human resources. To verify reliability, we randomly sampled 130 model outputs: GPT-5 assigned scores and provided rationales for each rubric subcategory. Out of 130 total samples, 115 (≈88.5%) were confirmed as accurate by an art expert, demonstrating strong agreement. Future work will expand human evaluations and include inter-rater reliability analysis. **Experiment Results:** As shown in Table 1, Gemini 2.5 Flash achieved the highest overall mean score (4.64), performing consistently across all analytic dimensions. Gemini 2.0 Flash performed best in Visual Analysis (4.84) but trailed in contextual reasoning, while GPT-4o-mini showed the strongest Cultural Significance (4.80). In contrast, GPT-o4-mini and GPT-4o scored lower, especially in historical and cultural dimensions. Overall, Gemini models demonstrated the most balanced and robust performance.

Table 1: Performance of LLMs based on GPT-5 rating scores across three analytic dimensions.

Model Name	Visual Analysis	Historical Context	Cultural Significance	Overall
Gemini 2.5 Flash	4.720	**4.600**	4.600	**4.640**
Gemini 2.0 Flash	**4.840**	4.325	4.525	4.563
GPT-4o-mini	4.400	4.425	**4.800**	4.542
GPT-o4-mini	4.740	3.600	3.433	3.924
GPT-4o	4.520	3.400	3.300	3.740

3 Web Application Development

As shown in Figure 1, we developed LLM4Art, a web-based application that visualizes model analyses. The system is built with a lightweight HTML and JavaScript frontend and a Node.js backend. The client resizes and encodes uploaded images locally before sending them to the backend API for analysis via LLMs. The modular design organizes outputs into three analytic dimensions: Visual Analysis, Historical Context, and Cultural Significance. The Visual Analysis module is further divided into subcategories such as color, line, style, shape, and artifact type for fine-grained interpretation. The system also produces structured JSON responses and supports flexible export options. Learners can explore side-by-side

interpretations from multiple LLMs to understand artworks across styles and historical periods. The backend currently integrates the best-performing Gemini 2.5 Flash API, but its modular design allows easy replacement with new LLMs as benchmarking results evolve. We will conduct user studies to evaluate the interface design for AI-assisted art education.

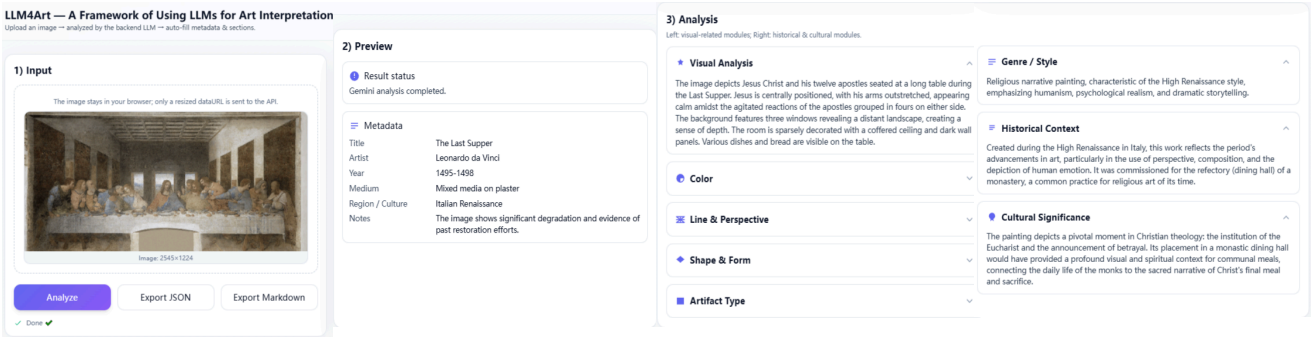


Figure 1: User interface of the LLM4Art application.

4 DISCUSSION AND FUTURE WORK

This study demonstrates that LLMs can feasibly interpret visual artworks in educational contexts when guided by co-designed prompts and rubrics. The results suggest that LLMs can produce art analyses, particularly for famous works. However, it remains unclear whether these models truly understand artistic elements or primarily rely on memorized knowledge from training data. Future work will expand the dataset to 5,000 diverse artworks, including both low-quality and high-quality artworks with more fine-grained human expert annotations. This larger dataset will support testing a broader range of LLMs, including open-source models, and enable more rigorous analyses of errors and bias. Human expert evaluation will also be scaled up to validate the results by incorporating inter-rater reliability measures to ensure fairness and transparency. Future work will focus on increasing expert involvement, testing open-source LLMs, and fine-tuning models for art-specific reasoning. We believe these efforts will enhance the reliability and educational value of LLM4Art as a tool for understanding and teaching art.

REFERENCES

Guidelines for Analysis of Art. (n.d.). School of Art and Design. Retrieved November 10, 2025, from https://ualr.edu/art/art-history-resources/guidelines-for-analysis-of-art/

Piper, A. (2016). *Art criticism: Suggested guidelines.* Retrieved November 10, 2025, from http://www.adrianpiper.com/docs/Piper2016ArtCriticismSuggestedGuidelines.pdf

Yuan, Z., He, Y., Wang, K., Ye, Y., & Sun, L. (2023). ArtGPT-4: Towards artistic-understanding large vision-language models with enhanced adapter. arXiv preprint arXiv:2305.07490.

Arita, T., Zheng, W., Suzuki, R., & Akiba, F. (2025). Assessing LLMs in Art Contexts: Critique Generation and Theory of Mind Evaluation. arXiv preprint arXiv:2504.12805.

Khadangi, A., Sartipi, A., Tchappi, I., & Fridgen, G. (2025). CognArtive: Large Language Models for Automating Art Analysis and Decoding Aesthetic Elements. arXiv preprint arXiv:2502.04353.

Bin, Y., Shi, W., Ding, Y., Hu, Z., Wang, Z., Yang, Y., ... & Shen, H. T. (2024, October). Gallerygpt: Analyzing paintings with large multimodal models. In Proceedings of the 32nd ACM International Conference on Multimedia (pp. 7734-7743).