

Generative AI and the Cross-Cultural Narration of Chinese Intangible Cultural Heritage: A Case Study of Shadow Puppetry

Zixuan Xu

Department of Artificial Intelligence, Macau University of Science and Technology, Macau, China

Abstract: Generative artificial intelligence is now frequently used to draft museum labels, tourism introductions, bilingual public texts, and short educational materials. In the field of intangible cultural heritage, however, a fluent paragraph can still miss the practical knowledge carried by performers, craftspeople, and local communities. Taking Chinese shadow puppetry as a case, this study examines how ChatGPT and ERNIE Bot narrate one living tradition for English-speaking readers. The analysis uses six heritage dimensions and four evaluation indicators: factual accuracy, cultural specificity, narrative accessibility, and commercialization risk. The comparison shows that both tools can produce readable first drafts and summarize basic facts, but their strongest passages often remain close to common public descriptions. More delicate elements, such as regional singing styles, hand manipulation behind the screen, apprenticeship, ritual occasions, and the economic pressure of tourism, are less consistently represented. The paper therefore argues for an AI-assisted, human-reviewed model of heritage communication, in which machine-generated drafts are treated as raw material rather than cultural authority.

Keywords: Generative AI; Intangible Cultural Heritage; Shadow Puppetry; Cultural Communication; AI Narrative

1. Introduction

Generative AI has entered many small tasks of cultural communication: writing short exhibition captions, turning Chinese introductions into English, polishing travel copy, and preparing background notes for international readers. Its usefulness is obvious because it reduces the cost of producing readable multilingual text. For Chinese traditional culture, this convenience seems particularly attractive, since many local practices need wider public explanation. Yet the

difficulty of cultural communication is not only whether a sentence is grammatical or whether a term is translated correctly. When a model explains a living tradition, it tends to choose details that are already frequent in public data, easy to generalize, or appealing to outside audiences. Research on large language models has warned that fluency may hide bias, simplification, and weak accountability [1]. For this reason, AI-written heritage narratives need to be read as interpretations, not as neutral containers of culture.

Intangible cultural heritage is sensitive because its core is not an object that can be fully described from the outside. UNESCO defines it through practices, representations, expressions, knowledge, and skills transmitted by communities and groups [2]. Chinese shadow puppetry, which entered UNESCO's Representative List in 2011, is a useful example because it combines carved leather figures, a lit screen, hand manipulation, music, local opera elements, oral narration, and public performance [3]. A puppet figure may be displayed in a museum case, but the tradition itself also depends on the voice of the performer, the timing of the musician, the memory of the script, and the social occasions in which people still watch the performance. If an AI text introduces it only as an ancient visual entertainment, the wording may be correct in a narrow sense while still leaving out what makes the practice living heritage.

The main question of this paper is how generative AI reconstructs Chinese shadow puppetry when it explains the tradition in English. A useful heritage narrative should be reliable in basic facts, specific enough to keep local texture, accessible enough for non-specialist readers, and cautious about turning culture into a saleable image. A weak narrative can still sound polished; it may use beautiful words while reducing heritage to a tourist attraction, an exotic symbol, or a simple national label. The tension between readability and

cultural depth is the reason this paper separates narrative accessibility from cultural specificity instead of treating them as the same achievement. This problem is specified through three research questions that are aligned with the three questions used in Section 3.1. RQ1 asks whether generative AI can explain the basic facts of Chinese shadow puppetry in English without obvious distortion. RQ2 asks whether it preserves culturally specific information, including regional performance settings, carving techniques, singing and narration, and living transmission. RQ3 asks whether it increases commercialization risk by turning the tradition into promotional language for tourism or cultural consumption. Together, these questions connect the article's opening problem with the later experimental design.

The article is organized around the chain of problem, experiment, result analysis, and conclusion. Section 2 reviews research on intangible cultural heritage, interpretation, and AI-generated cultural narratives. Section 3 explains the experimental setting, including the six shadow-puppetry dimensions, the two AI tools, and the scoring indicators. Section 4 discusses the results with three tables and three figures, paying special attention to the gap between smooth English expression and cultural adequacy. The conclusion returns to the practical question of how AI should be used in heritage communication: as a drafting assistant, not as a replacement for human cultural judgement.

2. Related Work

2.1 Intangible Cultural Heritage and Living Transmission

The concept of intangible cultural heritage shifts attention from preserved objects to practices that people continue to perform, teach, and reinterpret. The 2003 UNESCO Convention highlights community recognition, transmission, safeguarding, and respect for cultural diversity [2]. Heritage studies further remind us that heritage is not simply inherited from the past; it is selected and given meaning in the present [4]. Kirshenblatt-Gimblett describes intangible heritage as a form of metacultural production, because communities and institutions actively frame cultural practices as heritage [7]. This perspective matters for shadow puppetry. The carved figure is only one visible part of a larger practice that also includes leather preparation,

coloring, jointing, manipulation rods, singing, percussion, dialect-based narration, apprenticeship, and performance occasions. An English-language account that focuses only on puppets as beautiful artifacts therefore risks turning a living practice into a static museum object.

2.2 Heritage Interpretation and Cross-cultural Narration

Heritage interpretation is the process of making cultural practices understandable without stripping away their meaning. Tilden's classic argument is that interpretation should reveal significance rather than merely transfer information [5]. In cross-cultural narration, this principle becomes a practical problem of balance. A text that is overloaded with local terms, historical dates, and technical craft vocabulary may be hard for international readers to enter. A text that removes those details completely may become pleasant but hollow. Chinese shadow puppetry makes this balance especially visible because one performance may involve craft knowledge, local opera, folk belief, family training, public entertainment, and regional identity at the same time. Good interpretation therefore needs both plain explanation and selective depth.

2.3 Generative AI and Cultural Narrative Risk

Generative AI brings a new complication into heritage interpretation: it can produce confident, coherent cultural prose without having any lived relation to the community being described. Research on language models shows that they generate text from statistical patterns rather than from accountable participation in social life [1]. As a result, their heritage narratives often repeat dominant public descriptions while overlooking local voices, minority details, and contradictions inside the tradition. This is not a small stylistic issue. Heritage narratives shape cultural memory, influencing what later audiences remember, quote, and circulate [8]. The technical strength of large models can make the problem harder to notice, because fluent output may appear authoritative even when cultural detail is thin [9]. Digital heritage scholarship has already shown that technology can support preservation and dissemination while also detaching cultural expression from its original social setting [6]. For evaluation, it is useful to distinguish types

and seriousness of error, just as error levels can be compared in other AI language tasks [10]. This paper therefore assesses AI texts through both communication effectiveness and cultural responsibility.

3. Experimental Design

3.1 Research Questions and Task Setting

The experiment is built around the same three research questions introduced above. RQ1 asks whether generative AI can explain the basic facts of Chinese shadow puppetry in English without obvious distortion. RQ2 asks whether it keeps culturally specific information, such as regional performance settings, carving techniques, singing and narration, and living transmission. RQ3 asks whether it increases commercialization risk by turning the tradition into promotional language for tourism or cultural consumption. The task is therefore broader than translation. It asks how an AI system selects, orders, and tones a living cultural tradition for readers who may have little prior knowledge of China.

3.2 Case Dimensions and AI Systems

In this paper, the dataset refers to the prompt-output pairs used for analysis, rather than to the analytical dimensions themselves. The study uses six prompts, each corresponding to one dimension of Chinese shadow puppetry: origins and regional variation, leather carving and visual symbols, music and oral narration, ritual and festival context, apprenticeship and transmission,

and tourism and digital adaptation. ChatGPT and ERNIE Bot each generated one English output for every prompt, producing 12 model-generated English texts, or 12 prompt-output pairs, in total. The six dimensions function as analytical categories for comparing these texts. They were chosen because they move from what is easiest to see, such as puppet images, to what is easier to overlook, such as voice, training, occasion, and community use. The dataset is deliberately small, so the study does not claim statistical representativeness. Its value lies in making a qualitative comparison visible: the same tradition is repeatedly narrated under different prompts, and the scoring records where the narratives become strong, thin, or risky.

To make the experiment reproducible, the prompt protocol was kept identical across the two systems. No multi-turn dialogue was used. The system instruction was: "You are a cultural heritage communicator writing for English-speaking non-specialist readers. Explain Chinese shadow puppetry accurately, preserve local cultural detail, and avoid turning the tradition into tourism advertising." The user prompt template for each case was: "Write an English introduction of about 180 words on the following dimension of Chinese shadow puppetry: [dimension]. Include factual background, culturally specific details, and a balanced tone. Do not add unsupported claims or promotional language." Each dimension in Table 1 was inserted once into the template for ChatGPT and once for ERNIE Bot. The full prompt template is repeated in Appendix A.

Table 1. Case Dimensions and Prompt Focus Used in the Shadow-Puppetry Experiment

Case	Heritage dimension	Analytical focus	Reason for inclusion
1	Origins and regional variation	Historical spread, regional styles, local performance ecology	Tests whether AI avoids treating shadow puppetry as one uniform national form.
2	Leather carving and visual symbols	Materials, carving, coloring, puppet figures, symbolic patterns	Covers the most visible and easily described craft features.
3	Music and oral narration	Voice, singing style, rhythm, local opera elements, storytelling	Tests whether AI notices performance practices beyond the puppet image.
4	Ritual and festival context	Occasions, community use, festival performance, folk belief	Tests whether AI connects the art form to social and ritual settings.
5	Apprenticeship and transmission	Training, performer knowledge, family or community	Tests whether AI recognizes living transmission and practitioner
6	Tourism and digital adaptation	Tourism packaging, online circulation, modernization, commercialization	Tests whether AI balances contemporary communication with cultural responsibility.

3.3 Evaluation Indicators and Calculation Method

Four indicators are used: factual accuracy, cultural specificity, narrative accessibility, and commercialization risk. Each case receives a

score from 1 to 5 for each AI system. Factual accuracy concerns whether the text correctly describes the basic history, form, and practice of shadow puppetry. Cultural specificity records whether local context, regional difference, technical detail, and practitioner agency remain

visible. Narrative accessibility measures clarity, readability, and the ability to guide a non-specialist English reader through the subject. Commercialization risk is scored separately because attractive language is not always responsible language; it measures the extent to which the output packages heritage as spectacle, tourism, or a simplified cultural product. For the first three indicators, a higher score means stronger performance. For commercialization risk, a higher score means greater risk. The average score in Figure 1 is the arithmetic mean of the six case scores for each tool and indicator. Figure 2 reports the six cultural-specificity scores by case, and Figure 3 places accessibility and commercialization risk in the same view. Table 2 makes the scoring logic explicit so that the discussion does not rely only on subjective impression.

The decimal values in Table 3 are composite scores rather than separate scores from multiple evaluators. For each model, case, and indicator,

the author-evaluator applied ten equally weighted diagnostic checkpoints derived from the relevant rubric and assigned each checkpoint an integer value from 1 to 5. The reported case score is calculated as $S(m,c,i) = (r_1 + r_2 + \dots + r_{10}) / 10$, where m is the model, c is the case, i is the indicator, and r_1 - r_{10} are the checkpoint ratings. This procedure permits one-decimal increments, such as 3.2 or 3.6, while preserving the underlying 1-to-5 scale. The tool-level score is then calculated as $T(m,i) = (S(m,1,i) + S(m,2,i) + \dots + S(m,6,i)) / 6$. The evaluator was the author of the study, who has training in English-language academic writing and Chinese cultural-heritage reading. The scoring was completed with the same rubric for all outputs and then checked a second time after a short interval to reduce immediate impression bias. Because the present study uses a single evaluator and one output per model for each prompt, the scores should be read as transparent analytic ratings rather than as inter-rater reliability statistics.

Table 2. Evaluation Indicators and Scoring Rules

Indicator	Calculation and scoring rule	Interpretation
Factual accuracy	Ten 1-to-5 checkpoint ratings are averaged for each case, focusing on the correctness of basic history, form, and practice. The tool-level score is the mean of six case scores.	Higher score means better factual reliability.
Cultural specificity	Ten 1-to-5 checkpoint ratings are averaged for each case, focusing on local context, regional difference, practitioner agency, and technical detail. The tool-level score is the mean of six case scores.	Higher score means stronger preservation of cultural depth.
Narrative accessibility	Ten 1-to-5 checkpoint ratings are averaged for each case, focusing on clarity, readability, and engagement for non-specialist English readers. The tool-level score is the mean of six case scores.	Higher score means stronger reader accessibility.
Commercialization risk	Ten 1-to-5 checkpoint ratings are averaged for each case, focusing on whether the output packages heritage as spectacle, tourism, or a simplified cultural product. The tool-level score is the mean of six case scores.	Higher score means greater risk and weaker cultural responsibility.

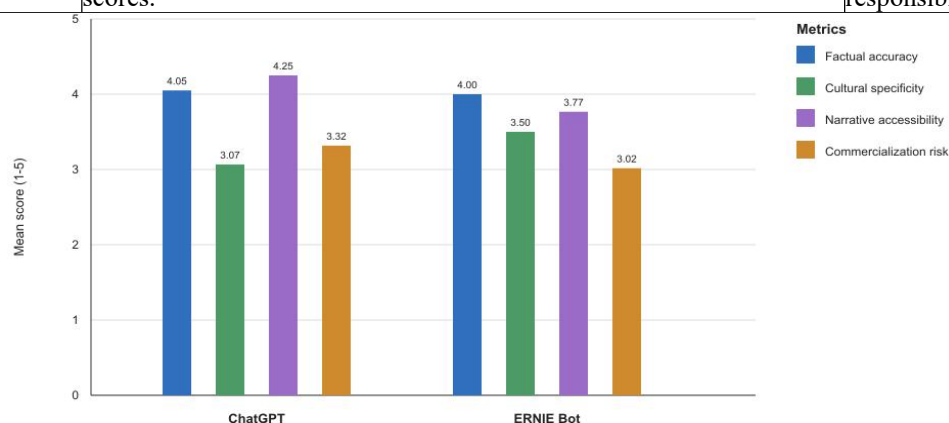


Figure 1. Average Evaluation Scores of AI-generated Shadow Puppetry Narratives.

4. Results and Analysis

4.1 Overall Performance Across Indicators

Figure 1 gives a first overview of the two

systems. Both tools are relatively strong in factual accuracy and narrative accessibility, which means that they can explain the basic form of shadow puppetry and produce English prose that is easy to follow. ChatGPT has the

higher accessibility score. Its paragraphs tend to move quickly from definition to historical background and then to cultural value, a structure that is convenient for readers who need a short introduction. ERNIE Bot shows a slightly higher cultural-specificity score, partly because it keeps more China-centered terms and gives more attention to heritage status. The more

important finding, however, is the distance between accessibility and cultural depth. A text can be clear and still be too smooth: it may explain the screen, light, and puppets while saying little about the performer behind the screen or the regional community around the performance.

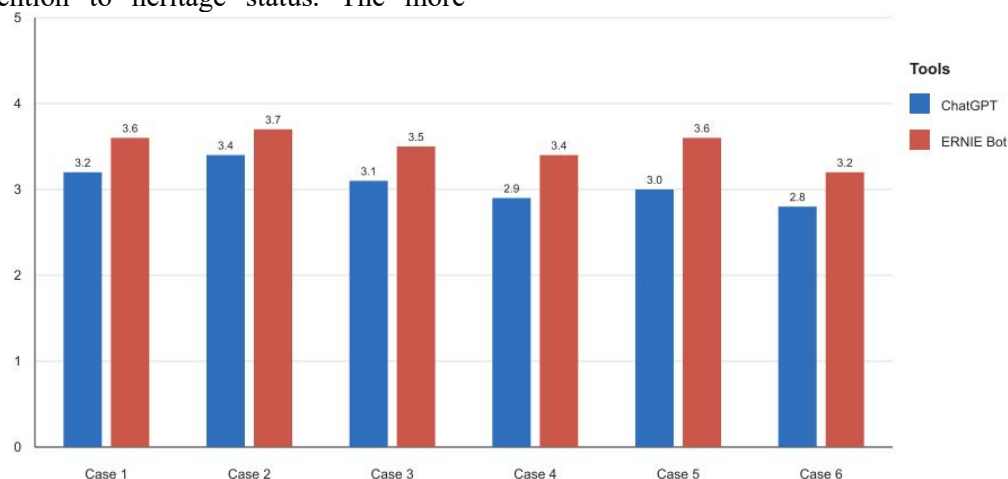


Figure 2. Cultural Specificity Scores Across Six Heritage Dimensions

Table 3. Full Case-level Scoring Matrix

Case	Tool	Factual accuracy	Cultural specificity	Narrative accessibility	Commercialization risk
1	ChatGPT	4.4	3.2	4.3	3.1
1	ERNIE Bot	4.2	3.6	4.0	2.9
2	ChatGPT	4.5	3.4	4.4	3.0
2	ERNIE Bot	4.3	3.7	4.0	2.8
3	ChatGPT	4.1	3.1	4.0	3.3
3	ERNIE Bot	4.0	3.5	3.8	3.0
4	ChatGPT	4.0	2.9	3.9	3.5
4	ERNIE Bot	4.0	3.4	3.7	3.1
5	ChatGPT	4.2	3.0	4.1	3.2
5	ERNIE Bot	4.1	3.6	3.8	2.9
6	ChatGPT	3.9	2.8	4.2	3.8
6	ERNIE Bot	3.8	3.2	3.9	3.4

Note. CS = cultural specificity; CR = commercialization risk. Scores range from 1 to 5.

4.2 Cultural Specificity Across Heritage Dimensions

Figure 2 shows that cultural specificity changes noticeably across the six dimensions. Both systems perform better when the topic is visible and concrete, especially leather carving, puppet figures, and visual symbols. This is understandable because such features are often described in public introductions and can be summarized with stable vocabulary. The scores become weaker when the prompt concerns oral narration, ritual and festival context, apprenticeship, and community transmission. These dimensions are harder for the models because they require attention to practice rather

than appearance. For example, a useful account of music and narration should mention that the story is carried by voice, rhythm, and local theatrical convention, not simply by the puppet image. Table 3 makes this pattern clearer. The tourism and digital adaptation case has the lowest cultural-specificity scores and the highest commercialization-risk scores, showing that AI texts become less culturally precise when they try to make the tradition more attractive and contemporary

4.3 Accessibility and Commercialization Risk

Figure 3 places narrative accessibility beside commercialization risk, and the relationship is not accidental. The most readable AI passages often rely on familiar promotional adjectives, such as ancient, colorful, magical, unique, or

must-see. These words help a reader form a quick impression, but they also push shadow puppetry toward the language of tourism advertising. In the tourism and digital adaptation case, ChatGPT in particular tends to describe the tradition as an attractive cultural experience rather than as a practice facing transmission

pressure. ERNIE Bot is slightly more restrained, but it also sometimes uses official-sounding heritage praise instead of describing concrete practice. This result suggests that AI-generated heritage text should be checked not only for factual mistakes but also for tone, emphasis, and the kind of audience relationship it creates.

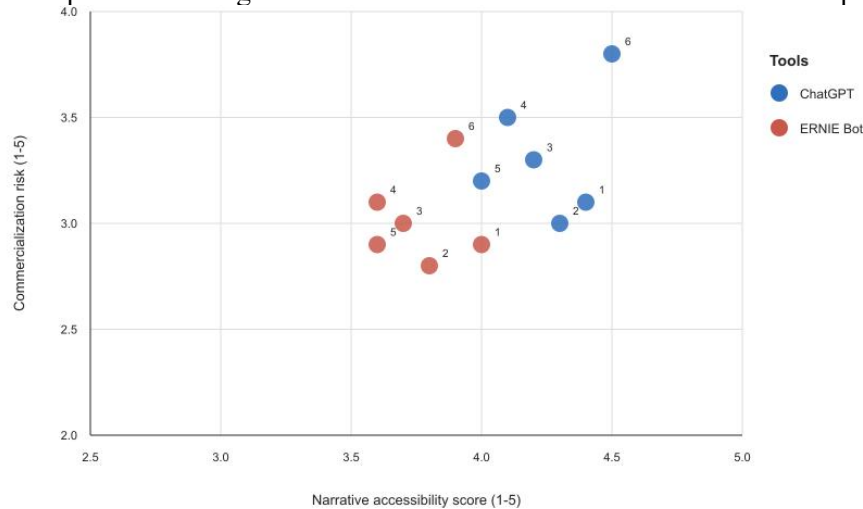


Figure 3. Narrative Accessibility and Commercialization Risk in AI Heritage Narratives.

4.4 Discussion of the Problem-experiment-result Chain

The experiment answers the central problem with a mixed result. Generative AI is useful at the entry level: it can draft an English explanation, organize basic background, and make an unfamiliar art form less distant for international readers. At the same time, the results show why this usefulness should not be confused with cultural authority. The models are better at the visible surface of shadow puppetry than at the social relations that sustain it. They also tend to solve the problem of readability by making the tradition attractive, which can quietly increase commercialization risk. A more responsible workflow would be AI-assisted heritage narration. In that workflow, AI supplies a draft or several alternative phrasings, while practitioners, researchers, translators, or curators add local context, remove promotional exaggeration, correct missing details, and decide what should not be simplified.

5. Conclusion

This paper examined how generative AI reconstructs Chinese shadow puppetry for English-speaking audiences. It began from a practical problem: AI-generated cultural narratives can be fluent, readable, and useful while still simplifying the living conditions of

intangible heritage. To test this problem, the study compared ChatGPT and ERNIE Bot across six dimensions of shadow puppetry and evaluated their outputs with four indicators. The results show that both systems can produce acceptable first-level explanations, especially when the task involves basic facts or a general introduction for non-specialist readers. Their limitations appear more clearly when the topic requires cultural specificity. Regional variation, oral performance, ritual use, apprenticeship, and practitioner agency are not always carried into the English narrative with enough detail. Commercialization risk is also a repeated concern because attractive language can make heritage look like a polished cultural product rather than a practice maintained through labor, memory, and community use. These conclusions should be understood within a limited scope. They apply only to the six shadow-puppetry prompt cases examined here and should not be generalized to all Chinese intangible cultural heritage, all heritage communication tasks, or the entire field of generative AI translation and narration. The study also relies on one author-evaluator, so future work should include multiple trained raters and report inter-rater reliability. Within this limited scope, the conclusion is not that generative AI should be rejected. A more realistic conclusion is that it should be used with a clear division of

responsibility: AI may draft, translate, summarize, and offer alternative wording, but human experts and cultural practitioners must review what is emphasized, omitted, or overpackaged. Future research could enlarge the dataset to other Chinese intangible heritage forms, such as paper cutting, Kunqu opera, calligraphy, or traditional festivals, and could include feedback from practitioners and international readers to test whether the scoring framework matches actual reception.

References

- [1] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623). ACM.
- [2] UNESCO. (2003). *Convention for the safeguarding of the intangible cultural heritage*. UNESCO.
- [3] UNESCO. (2011). *Chinese shadow puppetry. Representative List of the Intangible Cultural Heritage of Humanity*.
- [4] Smith, L. (2006). *Uses of heritage*. Routledge.
- [5] Tilden, F. (1977). *Interpreting our heritage* (3rd ed.). University of North Carolina Press.
- [6] Cameron, F., & Kenderdine, S. (Eds.). (2007). *Theorizing digital cultural heritage: A critical discourse*. MIT Press.
- [7] Kirshenblatt-Gimblett, B. (2004). *Intangible*

heritage as metacultural production. *Museum International*, 56(1-2), 52-65.

- [8] Nora, P. (1989). Between memory and history: Les lieux de memoire. *Representations*, 26, 7-24.
- [9] OpenAI. (2023). GPT-4 technical report. arXiv. <https://arxiv.org/abs/2303.08774>
- [10] Wang, Q., Ding, K., Gao, H., Wang, H., & Xu, R. (2025). Error comparison optimization for large language models on aspect-based sentiment analysis. Unpublished manuscript.

Appendix A. Prompt protocol

The following protocol was used for both ChatGPT and ERNIE Bot. System instruction: "You are a cultural heritage communicator writing for English-speaking non-specialist readers. Explain Chinese shadow puppetry accurately, preserve local cultural detail, and avoid turning the tradition into tourism advertising." User prompt template: "Write an English introduction of about 180 words on the following dimension of Chinese shadow puppetry: [dimension]. Include factual background, culturally specific details, and a balanced tone. Do not add unsupported claims or promotional language." Each dimension in Table 1 was inserted once into the template for each AI system. No additional follow-up prompts were used.