Can LLMs Help Sun Wukong in his Journey to the West?
A Case Study of Language Models in Video Game Localization

Xiaojing Zhao   Han Xu   Huacheng Song   Emmanuele Chersoni   Chu-Ren Huang
Department of Language Science and Technology, The Hong Kong Polytechnic University
{xiaojing.zhao, huacheng.song}@connect.polyu.hk
{han12.xu, emmanuele.chersoni, churen.huang}@polyu.edu.hk

Abstract

Large language models (LLMs) have demonstrated increasing proficiency in general-purpose translation, yet their effectiveness in creative domains such as game localization remains underexplored. This study focuses on the role of LLMs in game localization from both linguistic quality and sociocultural adequacy through a case study of the video game *Black Myth: Wukong*.

Results indicate that LLMs demonstrate adequate competence in accuracy and fluency, achieving performance comparable to human translators. However, limitations remain in the literal translation of culture-specific terms and offensive language. Human oversight is required to ensure nuanced cultural authenticity and sensitivity. Insights from human evaluations also suggest that current automatic metrics and the Multidimensional Quality Metrics framework may be inadequate for evaluating creative translation. Finally, varying human preferences in localization pose a learning ambiguity for LLMs to perform optimal translation strategies. The findings highlight the potential and shortcomings of LLMs to serve as collaborative tools in game localization workflows. Data are available at https://github.com/zcocozz/wukong-localization.

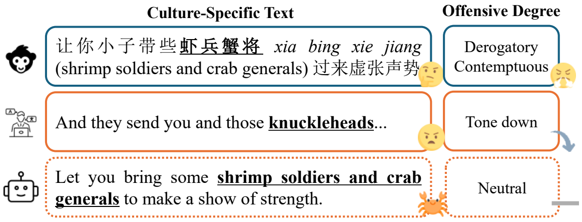

Culture-Specific Text
Offensive Degree
让你小子带些虾兵蟹将 *xia bing xie jiang* (shrimp soldiers and crab generals) 过来虚张声势
Derogatory Contemptuous
And they send you and those **knuckleheads**...
Tone down
Let you bring some **shrimp soldiers and crab generals** to make a show of strength.
Neutral

Figure 1: Human and LLM translations for cultural and offensive language in *Black Myth: Wukong*.

1 Introduction

Recent advances in large language models (LLMs) have significantly expanded the frontiers of machine translation (MT), achieving state-of-the-art performance across technical and literary domains (Hendy et al., 2023; Jiao et al., 2023; Wang et al., 2023). Unlike conventional MT systems, which struggle with idiomatic expressions and context-dependent scenarios, LLMs demonstrate potential in handling complex linguistic phenomena, including metaphor and idioms (Stowe et al., 2022; Tang et al., 2024; Yue et al., 2024). Moreover, recent studies suggest that LLMs can match the performance of junior human translators, signaling progress toward human parity in MT (Yan et al., 2024).

Despite these advancements, LLMs' potential for **video game localization** remains underexplored. The massive AAA game industry requires rapid translation of high-budget games from publishers into multiple languages for simultaneous global releases. While game localization teams face constraints including time and resources, the multilingual capabilities of LLMs offer a promising solution. Moreover, initiatives like Sony's China Hero Project aim to introduce more Chinese games into the global market[1], intensifying the need for culture-adapted multilingual translations. These games, rich in culture- and history-specific imagery, present a unique challenge of balancing preserving original cultural nuances with reshaping content to resonate with Western audiences.

This situation underscores a core dilemma in game localization—maintaining fidelity to the source material while adapting it appropriately for the target market. On one hand, successful localization requires preserving the "look and feel" of the original version by retaining key elements that define the game (O'Hagan and Mangiron, 2006). On the other hand, it demands sociocultural adaptation

[1] https://www.playstation.com/en-us/china-hero-project/

Proceedings of the First Workshop NLP and Language Models for Digital Humanities associated with RANLP 2025, pages 164–173, Varna, Bulgaria, Sep 11, 2025.
https://doi.org/10.26615/978-954-452-106-6-016

to meet target market expectations and avoid cultural sensitivities, with offensive language emerging as a prominent concern (Al-Batineh, 2021). This tension makes ensuring authentic gaming experiences across diverse cultural contexts inherently challenging. As illustrated in Figure 1, game localization needs to ensure sociocultural adequacy, including culture-specific terms and offensive language. Games embed cultural references, slang, and humor that demand context-aware translation. For instance, human translators adapt the mythological term “虾兵蟹将 *xia bing xie jiang*” into a colloquial equivalent “knuckleheads”, whereas LLMs produce literal translation like “shrimp soldiers and crab generals”, confusing audiences unfamiliar with the source culture. Such equivalents may also diminish original offensiveness by stripping culturally charged connotations.

To investigate the potential of LLMs for video game localization, this work presents a systematic evaluation combining automatic metrics with human assessments, using the recent *Black Myth: Wukong* game as a case study. We examine both the linguistic quality (accuracy, fluency) and sociocultural adequacy (cultural appropriateness, offensiveness rating) of LLM translations.

Our findings reveal mixed capabilities of LLMs in game localization. LLMs excel in linguistic quality, delivering satisfactory accuracy and fluency, yet they struggle with the cultural adaptation of culture-specific terms and offensive language. Human evaluations further suggest that automatic metrics and Multidimensional Quality Metrics (MQM) standards may not be appropriate for evaluating creative translation, while diverse human preferences in localization also pose learning ambiguity for LLMs to identify optimal translation strategies. The evolving capabilities of LLMs suggest their potential as collaborative partners in game localization workflows, though human post-editing remains essential for maintaining cultural authenticity and addressing cultural sensitivity. To our knowledge, this study represents the first systematic evaluation of LLMs in video game localization.

2 Related Work

2.1 LLMs for Translation

Recent studies demonstrate that LLMs can rival or even surpass traditional MT systems, achieving performance that nears human parity in basic tasks, such as GPT-4 showing ability competitive with commercial MT systems (Jiao et al., 2023). Furthermore, evaluations across 102 languages reveal steady improvements in high-resource languages, although challenges still remain in low-resource contexts (Hendy et al., 2023; Zhu et al., 2024).

Beyond general-purpose translation, LLMs have achieved notable advances in specialized professional and literary domains where traditional MT systems typically struggle. In legal translation, GPT-4 produces contextually accurate outputs comparable to human performance (Briva-Iglesias et al., 2024). This capability extends to culturally complex tasks, with LLMs successfully navigating context-dependent challenges including idiomatic expressions and poetry translation (Chen et al., 2024; Tang et al., 2024; Yao et al., 2024).

Building on this progress, LLMs present promising applications for video game localization. However, despite their great potential, research investigating LLM performance in video gaming localization remains limited. While Moreno García and Mangiron (2024) examined GPT-4’s translation of *Pokémon* terminology and found that the model could successfully implement creative translations, the scope and scale need to be enlarged to encompass a broader range of linguistic contexts. Meanwhile, human evaluation is essential to assess the fine-grained quality of LLM translations.

2.2 Game Localization

The emergence of LLMs has unveiled novel opportunities in video game localization. Localization projects typically operate under severe time and budget constraints that can undermine creative adaptation and player experience (O’Hagan and Chandler, 2016). These limitations create a pressing need for cost-effective and efficient translation solutions. The ongoing progress of LLMs’ contextual reasoning and cultural adaptation capabilities presents promising opportunities for improving efficiency and translation quality in localization.

The core of successful game localization is to deliver authentic player experiences. It involves adapting in-game texts, audio, and visual elements to match the target language and cultural context while preserving the narrative intent. Research reveals a strong player preference for localization that preserves original cultural elements rather than adapting them to local norms, as cultural authenticity is essential to player engagement and gaming experience (Costales, 2016; Ellefsen and Bernal-Merino, 2018; Khoshsaligheh et al., 2020; Wu and Chen, 2020). However, retaining culture- and history-specific elements, such as idioms, metaphors, and slang, remains a persistent challenge due to the limited translatability across sociolinguistic contexts.

Game localization also faces significant challenges in delivering culturally sensitive content, particularly in regions with stringent regulatory or sociocultural norms. For instance, Arabic-localized games frequently undergo systematic sanitization of profanity, nudity, and alcohol through omission, substitution, or euphemistic translation (Mahasneh and Abu Kishek, 2018; Al-Batineh, 2021). While these practices comply with censorship requirements and cultural expectations, they frequently lead to a loss of semantic or pragmatic nuances.

3 Methodology

Data We select the blockbuster video game *Black Myth: Wukong* as our data source due to its unique cultural and linguistic representativeness. The game is adapted from the 16th-century Chinese classic *Journey to the West*, blending poetic allusions, religious themes, and vernacular dialogue. Its dual mission—promoting traditional Chinese culture while advancing global gaming experience—creates salient localization challenges: preserving culturally embedded idioms and folklore while ensuring accessibility for international players. Given the limited familiarity of Western audiences with ancient Chinese cultures, the localization of such a product poses significant complexity.

Although this game supports official subtitles in 12 languages on the interfaces, its voice dialogues are available only in Chinese and English. Accordingly, we focus on Chinese-to-English translations. We transcribe subtitles from official cutscenes and publicly available videos, followed by manual proofreading of all content. The resulting parallel corpus comprises 2,259 sentence-pairs, capturing diverse narrative elements: main and side request dialogues, chapter-ending narratives rich in cultural metaphors, and song lyrics. Although the corpus of the in-game subtitles may be modest in size, *Black Myth: Wukong* is still unique as a Chinese AAA game with a special historical background, and its stylistic spectrum, spanning from story-based in-game banter to literary prose, makes it an ideal test case for investigating the culture-aware translation capabilities of LLMs.

Pipeline To benchmark LLM performance from coarse-grained to fine-grained perspectives, we regard the collected official human translations in English from the game developer as gold references. By comparing LLM outputs with human references, we conduct a two-stage evaluation on mainstream LLMs: 1) first, we employ multiple automatic metrics to evaluate diverse LLMs across strategically varied prompts using a randomly sampled subset from our *Black Myth: Wukong* parallel corpus. The top-performing LLM-prompt configurations identified in this phase serve as the basis for subsequent human evaluation, where 2) we compare the translations generated by the optimally prompted LLM against human gold references across multidimensional linguistic quality (accuracy, fluency) and sociocultural adequacy (audience appropriateness, offensiveness handling) based on subsets of sampled cases as well as manually extracted offensive instances. Further details for each evaluation stage are presented in the following.

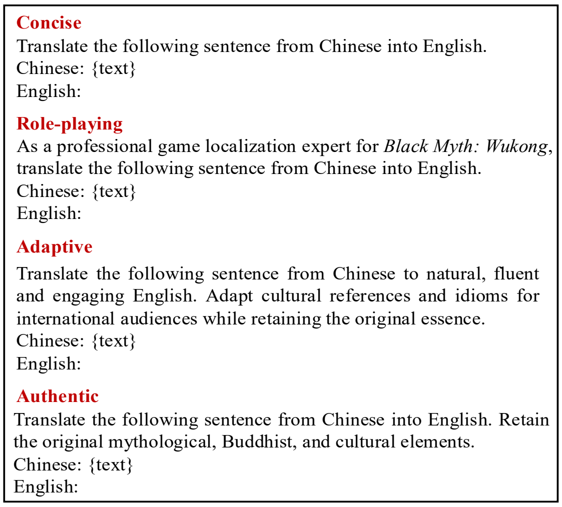
3.1 Automatic Evaluation

At this stage, we generate trial translations for 101 randomly sampled sentences from the complete corpus first using four open-source LLMs with four zero-shot prompting strategies, and then evaluate them with six automatic metrics, aiming to build a global view of LLMs' performance in localization and identify the optimal model-prompt pairing for subsequent human evaluation.

Models Our model selection comprises Llama3(-8B) (Grattafiori et al., 2024), TowerBase(-7B) (Alves et al., 2024), Qwen2(-7B) (Yang et al., 2024), and DeepSeek-LLM(-7B) (Bi et al., 2024), chosen for their known abilities in multilingual processing tasks. We choose all LLMs with around 7/8 billion parameters to compare the performance of different architectures at a similar parameter size.

Prompts Prior work in translation prompt engineering demonstrated that direct and minimalist prompts outperform complex formulations and achieve competitive performance (Jiao et al., 2023; Yan et al., 2024). In light of this, we develop four tailored and straightforward prompts addressing main localization requirements through different strategies: concise, role-playing, adaptive, and authentic, as detailed in Figure 2.

Metrics Using official English subtitles as gold-standard references, we automatically evaluate



Concise
Translate the following sentence from Chinese into English.
Chinese: {text}
English:

Role-playing
As a professional game localization expert for *Black Myth: Wukong*, translate the following sentence from Chinese into English.
Chinese: {text}
English:

Adaptive
Translate the following sentence from Chinese to natural, fluent and engaging English. Adapt cultural references and idioms for international audiences while retaining the original essence.
Chinese: {text}
English:

Authentic
Translate the following sentence from Chinese into English. Retain the original mythological, Buddhist, and cultural elements.
Chinese: {text}
English:

Figure 2: Different prompt types and specifications, where {text} represents the input source sentence.

models with six metrics. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020), COMET (Rei et al., 2022), XCOMET (Guerreiro et al., 2024), and XCOMET-QE (Guerreiro et al., 2024). They capture distinct performance aspects based on different principles: 1) string-overlap-based metrics (BLEU and ROUGE) emphasize surface-level equivalence; 2) among neural-network-based metrics (BERTScore, COMET, XCOMET, and XCOMET-QE), BERTScore uses the cosine similarity between token embeddings of candidate and reference texts, while the COMET-series metrics evaluate both meaning and form, as they are fine-tuned to predict human quality scores for translations. Among them, XCOMET-QE is an exception, as it assesses quality without any gold references by cross-lingually comparing source and target texts.

3.2 Human Evaluation

To further explore LLMs' culture-related capacity in game localization, we then conduct a human evaluation comparing the top-performing prompted LLM with gold human translations from aspects of linguistic quality in line with the MQM framework and sociocultural adequacy in terms of offensive language handling.

Multidimensional Translation Quality We evaluate translation quality using the MQM framework (Lommel et al., 2014), a comprehensive error typology enabling granular analysis of translation errors. Adopting the MQM core typology, we prioritize three key dimensions for localization: **accuracy** (semantic completeness and faithfulness), **fluency** (syntactic and grammatical correctness), and **audience appropriateness** (cross-cultural validity).

Offensive Language Annotation Beyond the MQM framework, we specifically assess sociocultural adequacy through offensive language handling. Offensive content was categorized along eight dimensions, spanning from explicit insult to culture-specific connotations. Two native Chinese speakers (C1 English-proficient video game players) independently labeled offensive expressions, achieving moderate agreement (Cohen's $\kappa = 0.48$). Discrepancies were resolved through consensus discussions with a third annotator, and 430 offensive cases were identified in the whole corpus.

Evaluation Protocol For both human evaluation tasks, ten postgraduate students specializing in Translation Studies were recruited as raters. They were paired into five groups, with each pair evaluating the same translations to ensure double annotation. Prior to the assessment, all these annotators received comprehensive training covering guidelines for the MQM translation error typology and offensive language classification, as well as gaining contextual familiarity with gameplay narratives through story walkthrough videos and detailed character biographies. Target human and machine translations were assessed on:

- **Translation Quality Scoring:** A five-point Likert scale for measuring accuracy, fluency, and audience appropriateness (from 1 = Poor to 5 = Excellent);
- **Offensive Language Rating:** A three-degree classification comparing the target translations to the source texts (less offensive, neutral, or more offensive).

To avoid the preference for human translations and biases to machine outputs, each rater conducted blind evaluations on a balanced mix of human and machine translations for sentences randomly selected from our original corpus. The evaluation yielded 1,972 ratings through a dual-rater process where each translation received independent ratings from two annotators. This included linguistic quality assessments of 900 translations across fluency, accuracy, and audience appropriateness, alongside offensiveness evaluations on 86 translations.

Although prior work has shown that MQM annotations typically achieve low inter-annotator agree-

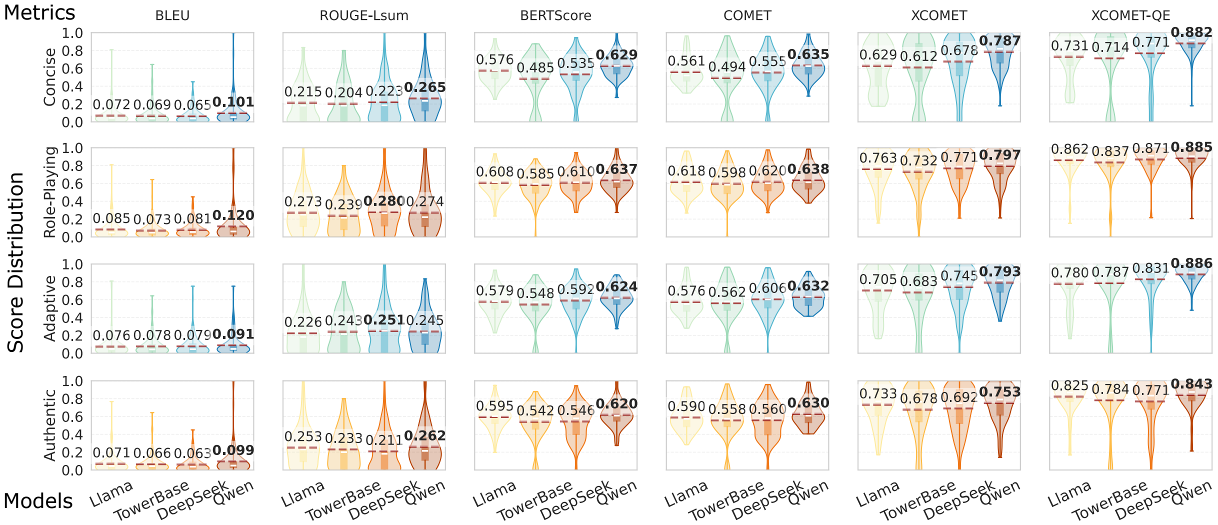


Figure 3: Model performance across different prompts at the sentence level. Each value represents the average score per LLM for the respective metric, with bold numbers highlighting the best performance within each prompt.

Task	Linguistics Quality		Sociocultural Adequacy	
Metric	*MQM Framework*			*Offensive Language*
Group	Accuracy	Fluency	Audience Appropriateness	Degree Rating
1	0.31	0.25	0.34	0.69
2	0.24	0.20	0.22	0.66
3	0.16	0.19	0.25	0.55
4	0.23	0.25	0.28	0.60
5	0.26	0.16	0.22	0.65

Table 1: Cohen's κ values for inter-rater agreement.

ment (Freitag et al., 2021), the framework remains valuable for identifying translation errors. As Table 1 shows, the Cohen's κ metric indicates a fair level of inter-annotator agreement overall for human evaluations in the present study. Notably, offensiveness ratings achieved higher consensus than translation quality assessment, as evaluating linguistic offensiveness is more straightforward than assessing holistic translation quality.

4 Results and Analysis

4.1 Automatic Evaluation

Results in Figure 3 show that LLMs achieve relatively low scores on n-gram overlap metrics, such as BLEU and ROUGE. In contrast, they achieve high scores on embedding-based metrics such as BERTScore and COMET-series, suggesting that LLMs make different lexical choices from reference translations, but still preserve source meaning.

In general, Qwen comes out on top among the four candidate models. When paired with role-playing prompts, it consistently delivers the best performance, only slightly trailing DeepSeek on the ROUGE metric. Therefore, we adopt Qwen with role-playing prompts as the exemplar LLM to generate the remaining translations for subsequent human evaluation.

4.2 Human Evaluation

This section presents human evaluation results from the MQM and offensive language ratings. Results indicate that while LLM translations achieve an adequate level of competence in linguistic quality comparable to human translations, they still fall short of sociocultural adequacy.

4.2.1 Linguistic Quality

Accuracy As shown in Table 2, LLM translations achieve scores comparable to human translations, with only marginal differences in accuracy. However, a closer analysis reveals that these similar scores hide different translation strategies. LLMs generally adopt a literal translation approach that prioritizes close alignment with the source text, yielding high accuracy scores when evaluations focus on textual correspondence. In contrast, human translations often incorporate creative adaptations to more effectively convey the underlying intent, which may result in deviations from the source text. For instance, while LLM translates "正好饿着 *zheng hao e zhe*" (happen to be hungry) as "I'm hungry", human translators render it as "Perfect timing", capturing the implied meaning that hunger coincides favorably with a meal opportunity. As a result, although human translations convey the communicative purposes, they may receive lower

Dimension	Group	Annotator	Human (M±SD)	LLM (M±SD)	t	p	Sig.
Accuracy	1	1	4.62±0.71	4.56±0.86	0.57	0.573	
		2	4.59±0.73	4.51±0.92	0.63	0.530	
	2	3	4.26±0.80	4.19±1.07	0.47	0.637	
		4	4.68±0.62	4.59±0.81	0.83	0.407	
	3	5	4.14±0.82	3.48±1.57	3.58	**< 0.001**	***
		6	4.32±0.78	4.62±0.88	-2.42	**0.016**	*
	4	7	4.73±0.65	4.70±0.76	0.32	0.752	
		8	4.11±1.08	4.66±0.81	-3.84	**< 0.001**	***
	5	9	4.16±1.03	4.01±1.30	0.83	0.410	
		10	3.82±1.41	3.49±1.67	1.45	0.150	

Table 2: Annotator's average accuracy ratings (M) with standard deviations (SD). Statistically significant differences (t, p, and Sig.) in scores between human and machine translations (Welch's t-test) are highlighted in bold.

scores under accuracy metrics emphasizing fidelity.

This observation suggests a potential limitation in applying MQM standards to creative texts. Unlike conventional translation tasks that prioritize accuracy in conveying propositional content, game localization often requires deliberate departures from the source text to achieve cultural adaptation, emotional resonance, and player engagement. Consequently, the emphasis on source-target fidelity may inadvertently penalize the creative translation that enhances quality in gaming contexts, indicating that specialized evaluation frameworks may be needed for assessing game localization.

Fluency Table 3 illustrates the model's capacity to generate fluent translations. Overall, the model performs strongly in terms of grammatical accuracy and punctuation. However, the lower fluency scores primarily stem from register inconsistencies and unnatural sentence flow, including awkward phrasing and redundant constructions that disrupt reading comprehension. These issues indicate persistent challenges in achieving stylistic precision for LLMs and directly affect player immersion in narrative-driven games. For instance, the LLM generates, "The evil monk *who incited* Jinchi Elder to *set fire to burn down* Tang Seng and his disciples *many years ago*". This translation suffers from redundant phrasing (*set fire to burn down*), and a passive, less engaging sentence structure caused by the use of a relative clause (*who*). In contrast, the human translation reads, "The evil monks *abetted* Elder Jinchi to *burn* the Great Sage and Tang Monk *alive*". This version is more concise and provides more vivid details (*burn ... alive*).

The difficulty of achieving natural fluency can be attributed to LLM's tendency toward literal translation, which results in rigid and lengthy sentences that disrupt the natural flow of dialogue. In contrast, human translators would restructure sentences to enhance readability. This observation aligns with previous findings that LLM outputs are generally more unnatural-sounding (Yan et al., 2024; Li et al., 2025). The linguistic complexity of our dataset also complicates the task, as it encompasses modern, classical, and vernacular Chinese variants. The diverse language styles require the model to balance formal linguistic structures with colloquial expressions, which creates a tension between maintaining structural fidelity and ensuring contextual fluency.

4.2.2 Sociocultural Adequacy

Audience Appropriateness Table 4 reveals a significant performance gap between LLM and human translations in tackling culture-specific terms, with LLM outputs consistently receiving statistically lower ratings. These relatively low ratings indicate that LLMs struggle with interpreting cultural nuances and appropriately conveying culture-specific terms, primarily due to their tendency toward cultural generalization and simplification. Table 5 illustrates culture-specific terms that require socioculturally aware translations. The Chinese expression "能耐 *neng nai*" (ability) carries nuanced connotations from genuine capability to sarcastic mockery, which require translators to interpret contextual cues to determine appropriate rendering. While human translators adapt their word choice to these varying contexts, LLM flattens this term to the emotionally neutral "ability" across all contexts and compromises the pragmatic information.

Beyond the issue of cultural neutralization that diminishes semantic connotations, the complexity of culturally appropriate translation is further compounded by diverse human preferences for localization strategies. The ratings for human translations exhibit notable variance and reflect disagreements about optimal cultural adaptation approaches. This

Dimension	Group	Annotator	Human (M±SD)	LLM (M±SD)	t	p	Sig.
Fluency	1	1	4.67±0.62	4.53±0.64	1.42	0.157	
		2	4.58±0.70	4.48±0.67	0.97	0.331	
	2	3	4.41±0.72	4.52±0.84	-0.96	0.341	
		4	4.84±0.47	4.74±0.73	1.09	0.276	
	3	5	4.78±0.68	3.23±1.79	7.64	**< 0.001**	***
		6	4.81±0.45	4.54±0.74	2.94	**0.004**	**
	4	7	4.79±0.59	4.62±0.73	1.69	0.093	
		8	4.59±0.78	4.51±0.82	0.65	0.516	
	5	9	4.57±0.72	4.42±1.02	1.10	0.273	
		10	4.20±1.32	3.83±1.44	1.78	0.076	

Table 3: Annotator’s average fluency ratings (M) with standard deviations (SD).

Dimension	Group	Annotator	Human (M±SD)	LLM (M±SD)	t	p	Sig.
Audience Appropriateness	1	1	4.78±0.54	4.39±0.98	3.30	**0.001**	**
		2	4.60±0.73	4.31±0.99	2.23	**0.027**	*
	2	3	4.41±0.73	4.36±1.01	0.42	0.673	
		4	4.82±0.41	4.48±0.85	2.46	**0.015**	*
	3	5	4.39±0.83	4.20±1.30	1.16	0.247	
		6	4.94±0.23	4.71±0.71	2.98	**0.004**	**
	4	7	4.61±0.86	4.08±1.25	3.34	**< 0.001**	***
		8	4.51±0.97	4.27±1.26	1.46	0.148	
	5	9	4.32±0.89	4.31±1.07	0.08	0.940	
		10	4.52±0.97	3.68±1.34	4.84	**< 0.001**	***

Table 4: Annotator’s average audience appropriateness ratings (M) with standard deviations (SD).

Source	这般能耐……正好正好
Human	A **strong foe**... Just what I need.
LLM	This kind of **ability**... just right, just right.
Source	有能耐，就在此间报仇罢！
Human	Then **if you can**, avenge him here and now!
LLM	If you have the **ability**, come and take revenge here!

Table 5: Contextual variations in translating the culture-specific term “能耐”.

complexity is exemplified in the translation of “妖怪 *yaoguai*” (monster), where annotators disagreed on the optimal strategy. While some favor direct transliteration as “yaoguai” to preserve cultural authenticity and introduce players to Chinese mythological concepts, others advocate for the culturally adapted equivalent “monster” to enhance immediate comprehension and gameplay accessibility.

The divergent perspectives highlight a fundamental tension in game localization: the competing demands of cultural preservation versus target audience accessibility in creating engaging player experiences. As human translators show divided preferences for translation strategies, LLMs face an ambiguous learning environment that lacks clear optimization targets. Consequently, game localization may necessitate continued human oversight for cultural appropriateness. A potential solution is to adopt a perspectivist approach to dataset creation, which involves capturing multiple preferences from different annotators instead of enforcing a single ground truth (Cabitza et al., 2023). This would enable LLMs to generate a spectrum of choices to better serve the nuanced demands of localization.

Offensiveness Handling Figure 4 illustrates the shifts of offensive intensity in human and LLM translations compared to the source texts. Human translations exhibit a broader range of variation, with more instances of increased and decreased offensiveness, which reflects strategic adjustments on a case-by-case basis. For mitigation, translators often substitute or omit highly offensive language, particularly expressions targeting religion or gender, to avoid offending and alienating the target audience. Occasionally, they amplify provocative elements for better characterization, plot development, and emotional resonance to strengthen narrative engagement and player immersion.

In contrast, LLM translations consistently maintain a neutral stance due to their reliance on literal translation that fails to account for cultural nuances. Since offensive language is deeply embedded in cultural context and often carries implicit meaning, LLMs struggle to capture the subtleties required for effective cross-cultural adaptation. This limitation is more pronounced with culture-specific

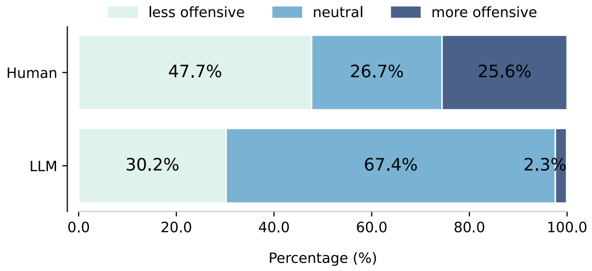
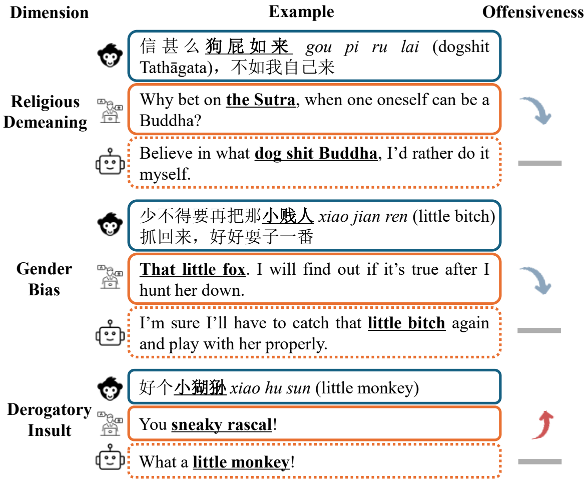


Figure 4: Distribution of offensive language intensity in human and LLM translations relative to source text.

expressions, such as idioms that demand cultural interpretation over direct linguistic conversion.

As exemplified in Figure 5, LLMs tend to preserve the original tone through literal translation, while human translators strategically adjust offensive language intensity to enhance immersive gameplay or align with target cultural norms. Interestingly, LLM's literal preservation can occasionally prove advantageous in gaming contexts where maintaining authentic character voices is crucial for narrative integrity, as such direct translation effectively conveys character traits like arrogance and rudeness. However, given localization requirements, human post-editing is needed to ensure cultural sensitivity in a global context.



Dimension	Example	Offensiveness
	信甚么**狗屁如来** *gou pi ru lai* (dogshit Tathāgata)，不如我自己来	
Religious Demeaning	Why bet on **the Sutra**, when one oneself can be a Buddha?	↘
	Believe in what **dog shit Buddha**, I'd rather do it myself.	—
	少不得要再把那**小贱人** *xiao jian ren* (little bitch) 抓回来，好好耍子一番	
Gender Bias	**That little fox**. I will find out if it's true after I hunt her down.	↘
	I'm sure I'll have to catch that **little bitch** again and play with her properly.	—
	好个**小猢狲** *xiao hu sun* (little monkey)	
Derogatory Insult	You **sneaky rascal**!	↑
	What a **little monkey**!	—

Figure 5: Examples of offensive language handling.

5 Conclusion

In this paper, we evaluate the performance of LLMs in game localization by examining linguistic quality and sociocultural adequacy. Our findings indicate that while LLMs demonstrate sufficient competence in general accuracy and fluency, they encounter challenges in contextual adaptation of culture-specific terms and offensive language. Evidence from human evaluations suggests that most automatic metrics and MQM standards may not be appropriate for evaluating creative translation. In addition, diverse human preferences in localization create learning ambiguity for LLMs to identify optimal translation strategies.

We consider our study as the first step to deepen our understanding of the strengths and limitations of LLMs in the translation of creative textual domains. As LLMs continue to evolve, our results highlight their promising potential as collaborative tools in the professional game localization workflow. Particularly in the context of the translation from Chinese to Western languages, where the biggest challenge is conveying meanings from an ancient culture with which those audiences are not familiar, we sincerely hope that LLMs can help Sun Wukong and the other heroes from future Chinese releases in their own Journey to the West.

Limitations

Our experiment exclusively employs open-source LLMs for reproducibility, thereby excluding proprietary systems such as the GPT series. Future studies could incorporate these models with larger sizes to enable comprehensive benchmarking, and clarify whether the limitations described above can be overcome through the use of larger-scale, more powerful architecture or they are inherent of the LLM paradigm. While our corpus provides creative contexts, its single-game focus limits the exposure to a wider range of linguistic patterns, potentially constraining the generalizability of our conclusions. This constraint highlights the pioneering and scarce availability of such culturally dense corpora. Additionally, although our Chinese-to-English focus aligns with the commercial demand for localizing Chinese games into global markets, game localization pipelines should support a larger number of languages and ensure fair treatment of all the cultures. Finally, due to the high cost of human annotation, we could evaluate only a randomly sampled subset of data. Our future work will address this limitation through more efficient annotation.

Acknowledgments

We express our sincere gratitude to the anonymous reviewers for their invaluable time and constructive suggestions. We also thank the annotators for their efforts and thorough work.

References

Mohammed Al-Batineh. 2021. Issues in arabic video game localization: A descriptive study. *Translation & Interpreting: The International Journal of Translation and Interpreting Research*, 13(2):45–64.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. 2024. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Vicent Briva-Iglesias, Joao Lucas Cavalheiro Camargo, and Gokhan Dogru. 2024. Large language models" ad referendum": How good are they at machine translation in the legal domain? *arXiv preprint arXiv:2402.07681*.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Benchmarking llms for translating classical chinese poetry: Evaluating adequacy, fluency, and elegance. *arXiv preprint arXiv:2408.09945*.

Alberto Fernández Costales. 2016. Analyzing players' perceptions on the translation of video games: Assessing the tension between the local and the global concerning language use. In *Media Across Borders*, pages 183–201. Routledge.

Ugo Ellefsen and Miguel Á Bernal-Merino. 2018. Harnessing the roar of the crowd: A quantitative study of language preferences in video games of french players of the northern hemisphere. *The Journal of Internationalization and Localization*, 5(1):21–48.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Masood Khoshsaligheh, Saeed Ameri, Farzaneh Shokoohmand, and Mehdi Mehdizadkhani. 2020. Subtitling in the iranian mediascape: Towards a culture-specific typology. *International Journal of Society, Culture & Language*, 8(2):55–74.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025. Lost in literalism: How supervised training shapes translationese in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12875–12894, Vienna, Austria. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.

Anjad A Mahasneh and Maysa'Taher Abu Kishek. 2018. Arabic localization of video games "tomb raider™(2013)": A start or a failure. *Lebende Sprachen*, 63(1):47–62.

Luis Damián Moreno García and Carme Mangiron. 2024. Exploring the potential of gpt-4 as an interactive transcreation assistant in game localisation: A case study on the translation of pokémon names. *Perspectives*, pages 1–18.

Minako O'Hagan and Carmen Mangiron. 2006. Game localisation: Unleashing imagination with" restricted" translation. *The Journal of Specialised Translation*, (6):10–21.

Minako O'Hagan and Heather Chandler. 2016. Game localization research and translation studies: Loss and gain under an interdisciplinary lens. In *Border Crossings*, pages 309–330. John Benjamins Publishing Company.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Kenan Tang, Peiyang Song, Yao Qin, and Xifeng Yan. 2024. Creative and context-aware translation of East Asian idioms with GPT-4. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9285–9305, Miami, Florida, USA. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Zhiwei Wu and Zhuojia Chen. 2020. Localizing chinese games for southeast asian markets: A multidimensional perspective. *The Journal of Internationalization and Localization*, 7(1-2):49–68.

Jianhao Yan, Pingchuan Yan, Yulong Chen, Judy Li, Xianchao Zhu, and Yue Zhang. 2024. Gpt-4 vs. human translators: A comprehensive evaluation of translation quality across languages, domains, and expertise levels. *arXiv preprint arXiv:2407.03658*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Shisen Yue, Siyuan Song, Xinyuan Cheng, and Hai Hu. 2024. Do large language models understand conversational implicature- a case study with a Chinese sitcom. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1270–1285, Taiyuan, China. Chinese Information Processing Society of China.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.