

Exploring Bengali Creative Storytelling Capabilities of Large Language Models Across Cultural Variations

Azmine Toushik Wasi
Shahjalal University of Science and
Technology
Sylhet, Bangladesh
azmine32@student.sust.edu

Raima Islam*
Harvard University
Cambridge, USA
raima_islam@g.harvard.edu

Mst Rafia Islam*
Independent University
Dhaka, Bangladesh
2030391@iub.edu.bd

Farig Sadeque
BRAC University
Dhaka, Bangladesh
farig.sadeque@bracu.ac.bd

Taki Hasan Rafi
Hanyang University
Seoul, South Korea
takihr@hanyang.ac.kr

Dong-Kyu Chae[†]
Hanyang University
Seoul, South Korea
dongkyu@hanyang.ac.kr

Abstract

Large Language Models (LLMs) excel in fluency but often struggle with originality, suspense, and emotional depth in storytelling. This study evaluates their creative storytelling capabilities in Bengali, a language with significant dialectal diversity. Using three narrative prompts across single-dialect and cross-dialect settings with initial results and story continuation, we analyze AI-generated content for coherence, creativity, and cultural relevance. Native Bengali speakers provide qualitative feedback, highlighting key challenges such as dialectal fidelity and narrative richness. Our findings emphasize the need for culturally adaptive NLP models to enhance storytelling in low-resource languages.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI); Interaction design.**

Keywords

Bias and Fairness, Bangla Storytelling, Dialects, LLM Assisted Writing, LLM Evaluation, LLM Auditing

ACM Reference Format:

Azmine Toushik Wasi, Raima Islam, Mst Rafia Islam, Farig Sadeque, Taki Hasan Rafi, and Dong-Kyu Chae. 2025. Exploring Bengali Creative Storytelling Capabilities of Large Language Models Across Cultural Variations. In *Proceedings of 2025 Conference on Human Factors in Computing Systems (CHI 2025)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/XX.XX>

*Both authors contributed equally to this research.

[†]Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI 2025, April 26 – May 1, 2025

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN XX

<https://doi.org/XX.XX>

1 Introduction

Large Language Models (LLMs) have shown impressive skills in generating coherent and grammatically sound narratives, yet their creative storytelling often lacks originality, suspense, and emotional depth. Research indicates that models such as GPT-4 and BART Large excel in fluency but tend to produce predictable and homogeneous stories [8, 9]. This limitation is partly due to their reliance on vast training datasets that can lead to overfitting on common narrative patterns. Benchmarks like CS4 reveal that these models struggle when tasked with highly specific creative constraints [1]. Concerns have also been raised about narrative ownership and the authenticity of AI-generated content. Recent innovations, such as Story3D-Agent, have attempted to enhance storytelling through procedural modeling and logical reasoning [6]. Despite these advancements, the core challenge of replicating the nuanced creativity found in human storytelling remains unresolved. As researchers explore the boundaries of creative AI, it is important to understand both the strengths and the shortcomings of these systems. This discussion sets the stage for a deeper evaluation of creative storytelling in diverse linguistic contexts.

Bengali, spoken by over 270 million people, offers a unique case study for evaluating AI-driven creative storytelling. The language is characterized by its dialectal diversity, with two major variants: the West Bengal dialect and the Bangladesh dialect. Although both dialects share a common linguistic foundation, subtle differences in vocabulary, expressions, and cultural connotations can significantly impact narrative style. For example, the word for “water” is expressed as ‘*jol*’ in the West Bengal dialect, while the Bangladesh dialect prefers ‘*pani*’ [4]. Such variations highlight the complexity of achieving linguistic authenticity in AI-generated content. Most existing LLMs are primarily trained on high-resource languages like English, which often results in a failure to capture these nuanced dialect differences. This shortcoming can lead to translations and creative outputs that lack cultural relevance. In the context of human-AI collaborative storytelling, maintaining cultural sensitivity while preserving creative freedom is especially crucial. Addressing these dialect-based challenges is essential to improve the overall user experience and mitigate potential biases in narrative generation.

In this study, we conduct a comprehensive evaluation of the creative storytelling capabilities of LLMs in Bengali, focusing specifically on dialectal variations. We designed three narrative prompts, each evaluated at two stages: story initialization and continuation. These prompts include two single-dialect examples and one cross-dialect example, ensuring a diverse representation of storytelling styles. To assess the effectiveness of AI-generated narratives, we collected qualitative feedback from native Bengali speakers, analyzing their perspectives on coherence, creativity, and cultural relevance. Our findings provide insights into the strengths and limitations of LLMs in Bengali storytelling, shedding light on key challenges such as dialectal fidelity, narrative richness, and user perception. This research contributes to the broader discourse on multilingual AI, highlighting the importance of culturally adaptive NLP models that can support creative expression in low-resource languages.

2 Background

Language and culture are deeply intertwined, and dialectal variations in Bengali reflect long-standing socio-cultural influences. Historical factors, including British colonial policies, have played a significant role in shaping the evolution of Bangla, contributing to the distinct characteristics of the West Bengal and Bangladesh dialects [4, 5, 10, 11]. These dialects are more than just different word choices; they serve as implicit markers of cultural and religious identity. Bengali Muslims in Bangladesh and Bengali Hindus in India exhibit distinct linguistic patterns that have been influenced by historical and cultural developments [3, 7]. As LLMs become increasingly used as writing tools, ensuring that they generate neutral and culturally sensitive outputs is paramount [3]. The challenge lies in balancing the creative assistance provided by these models with the need to respect and represent diverse cultural narratives. Our study aims to evaluate how well current LLMs can navigate these dialectal variations while supporting creative storytelling. By investigating these issues, we hope to contribute to the development of more inclusive and contextually aware AI systems [10, 11].

3 Methodology

We do a user satisfaction survey to ensure the dataset’s comprehensiveness and relevance while maintaining the study’s validity. This approach allows us to evaluate how effectively LLMs handle cultural nuances in language, ensuring they can engage respectfully and accurately with users from diverse backgrounds, specifically by experimenting with different dialect setups in user-centric creative writing. The user study includes names and evaluates both same-dialect and cross-dialect conversations in creative writing scenarios. It addresses two key aspects: initial LLM responses and their progression over continued interactions. We examine these interactions for both dialects to identify any existing biases. This study builds on our previous LLM-based evaluations by incorporating more controversial issues, such as religious words, names, and conversational complexities. These aspects could offer additional insights, as they are challenging to measure with standard evaluation methods. The user study will help us uncover and address these complex biases.

Setup. We have conducted a user study to thoroughly evaluate user satisfaction and understanding in writing assistance when the

LLM comprehends cultural complexities in dialects. We included 77 native Bangla-speaking participants who actively use LLMs, consisting of 42 males and 35 females, with most participants aged between 18 and 30. Regarding their LLM usage, the majority use it for academic purposes (61 participants out of 77), followed by creative uses (50 out of 77) and professional use (47 out of 77). Most participants are regular users, as shown in Figure 4.

Case Design and Development. We developed three example case stories using ChatGPT (GPT-4o) (as it performed the best compared to the other models), where the LLM successfully generated stories in Bangla. These stories included both dialects and cross-dialect usage. Each case consisted of two parts: *initialization and continuation*. Figure 2 presents all the cases, the LLM responses, and their English translations. Case 1 is mainly focused on West Bengal (Hindu majority) dialect, case 2 focuses on both dialects in a cross-lingual fashion, and case 3 is mainly focused on Bangladesh (Muslim majority) dialect. These cases were also verified by experienced native speakers and researchers. The translations of the case stories were made by ChatGPT and later verified and edited as needed by the authors.

Questionnaire Design. After presenting and evaluating the three cases, participants were asked to evaluate the entire experience through four statements: (3) *context awareness*, (4) *dialect and cultural sensitivity*, (5) *user retention*, (6) *consistency*, and provide one (7) *overall opinion*. We also collected a (8) *pre-evaluation (current) LLM usage satisfaction* score. All questions use a five-point scale, with answers ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). The questionnaire includes these questions:

- (1) **Correctness:** In this part, I thought the LLM worked well and correctly understood the dialect and related complexity.
- (2) **Quality:** In this part, I thought the LLM can be more helpful and provide better outcomes.
- (3) **Context Awareness:** I thought the system maintained context effectively throughout the writing process.
- (4) **Dialect and Cultural Sensitivity:** I thought the system understood and respected dialectal and cultural nuances in language.
- (5) **User Retention:** I would like to use LLMs for my future writing projects.
- (6) **Consistency:** I thought the system is consistent and works properly in different use cases.
- (7) **Overall Opinion:** I was pretty satisfied with the outcomes.
- (8) **Pre-evaluation (Current) LLM Usage Satisfaction:** I am satisfied with my current interactions with LLMs.¹

We designed the questions based on the use of LLM-based writing assistants in Bangla creative works and the System Usability Scale (SUS) [2], widely used in HCI for evaluating computing systems. The first two questions are repeated for each part, resulting in a total of 12 questions for 6 parts (2 for each; initialization and continuation) across 3 cases. For each part, participants rated their opinions on two aspects: (1) *correctness* and (2) *quality*. Higher value in (2) *quality* denotes current quality is lower and better quality is expected. A snapshot of the survey form is added in Figure 3.

¹It was asked before showing any case or providing any information.

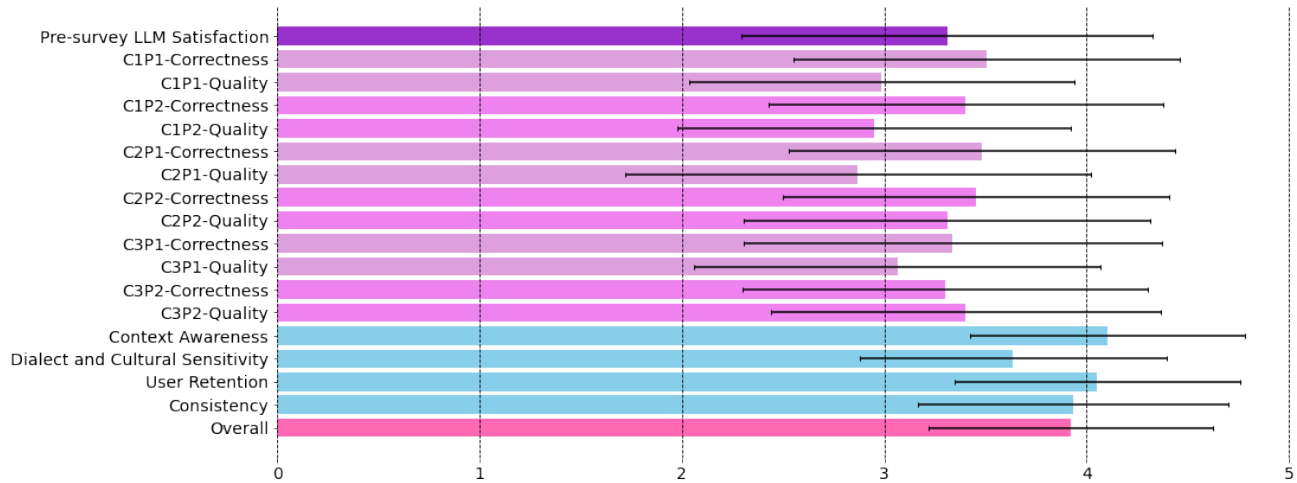


Figure 1: Detailed user study results.

4 Findings and Discussions

Figure 1 presents the results for the survey questions, where the length of each bar reflects the average score provided by the 77 participants, and the black line indicates the standard deviation of response values. The first 12 bars in Figure 1 show average values for each parts, in each questions (C1P1 means Case 1 Part 1: initialization; C1P2 means Case 1 Part 2: continuation, and so on).

4.1 Overall Findings and Analysis

The pre-survey satisfaction score was moderate at 3.31, with significant improvements seen post-survey, reaching an overall score of 3.92. Throughout the study, correctness consistently scored higher than quality across all cases, indicating that while the LLM generates accurate information, its presentation quality requires enhancement. Specifically, correctness scores ranged from 3.30 to 3.51, whereas quality scores varied more widely, from 2.87 to 3.40.

The analysis of the data reveals a consistent pattern where the correctness and quality scores for the continuation parts (P2) are generally lower than those for the initialization parts (P1) across most cases. For instance, in Case 1, Part 1 (C1P1), the correctness score is 3.51, which drops slightly to 3.40 in Part 2 (C1P2). Similarly, the quality score in C1P1 is 2.99, which decreases to 2.95 in C1P2. This trend is evident in Case 2 as well, where the correctness score declines from 3.48 (C2P1) to 3.45 (C2P2) and the quality score from 2.87 to 3.31, indicating some decrement in quality but still highlighting issues in continuation. Case 3 follows a similar pattern, with a decrease in correctness from 3.34 (C3P1) to 3.30 (C3P2), although there is a slight increase in quality from 3.06 to 3.40. These findings suggest that users perceive the LLM's continuation responses as more problematic and of lower quality compared to the initial responses, pointing to a need for enhancing the LLM's ability to maintain consistency and quality in extended interactions.

Context awareness received the highest mean score of 4.10, reflecting the LLM's strong ability to maintain and understand context, which is a critical strength. However, dialect and cultural sensitivity scored relatively lower at 3.64, suggesting a need for

incorporating more diverse dialects and cultural contexts into the training data. User retention and consistency also scored well, with means of 4.05 and 3.94, respectively, indicating that users find the LLM reliable and consistent. These findings suggest that while the LLM is effective and contextually aware, enhancing the quality of its outputs and its sensitivity to cultural nuances will further boost user satisfaction and overall effectiveness. By focusing on these improvement areas, the LLM can become a more valuable tool for its users, particularly in accurately and effectively handling diverse linguistic and cultural contexts.

4.2 Case-wise Analysis

Figure 1 provides a comparative analysis of the LLM's performance across three cases, focusing on different dialects and contexts.

Case 1 (West Bengal majority Dialect): The correctness scores show a slight decrease from Part 1 (C1P1: 3.51) to Part 2 (C1P2: 3.40), while the quality scores drop from 2.99 to 2.95. This indicates that the LLM maintains relatively stable performance in terms of correctness but shows a minor decline in quality during continuation. The standard deviations are quite high, suggesting variability in user perceptions.

Case 2 (Cross-Dialect): This case demonstrates a more pronounced change. The correctness scores decrease from 3.48 (C2P1) to 3.45 (C2P2), and the quality scores show a notable increase from 2.87 to 3.31. This suggests that while the LLM's correctness decreases slightly, the perceived quality of responses improves in continuation, potentially indicating better handling of cross-dialect interactions or an adaptation over time. The higher standard deviation in quality scores reflects greater variability in user feedback.

Case 3 (Bangladesh majority Dialect): Here, correctness scores decrease from 3.34 (C3P1) to 3.30 (C3P2), and quality scores increase from 3.06 to 3.40. The pattern in this case is similar to Case 2, with a slight decline in correctness but an improvement in quality over continuation. This improvement in quality might suggest that the LLM adapts well to the Bangladesh dialect over time.

Comparing Initialization and Continuation: Comparing Part 1 (P1) and Part 2 (P2) across the cases reveals a consistent trend

User Satisfaction Study : Case 1	
	<p>ছোট্ট অদिति আজ খুব সেজেছে, দুর্গাপূজা বলে কথা! সকাল থেকেই বাবার কাছে এসে বায়না, "বাবা, কখন মন্ডপে নিয়ে যাবে?", "বাবা চলানো" বাবারও সময় লাগছে কাজ শেষে বেরোতে। Little Aditi is all dressed up today, after all, it's Durga Puja! Since morning, she's been pestering her father, "Dad, when will you take me to the pandal?", "Dad, let's go!" Her father is also taking some time to finish his work before heading out.</p>
Part 1	<p>Initialization:</p> <p>বাবা: অদिति, একটু ধৈর্য ধরো মা, আমি একটু কাজ শেষ করে নিচ্ছি। তারপর আমরা মন্ডপে যাব। অদिति: কিন্তু বাবা, আমি তো অনেকক্ষণ ধরে অপেক্ষা করছি! বাবা: আমি জানি মা, আর মাত্র দশ মিনিট। তুমি ততক্ষণে তোমার মা'কে গিয়ে দেখাও তো, তুমি কেমন সুন্দর করে সেজেছ। Father: Aditi, please be patient, dear. I need to finish some work. After that, we will go to the pandal. Aditi: But Dad, I've been waiting for a long time! Father: I know, dear, just ten more minutes. In the meantime, why don't you go and show your mom how beautifully you've dressed!</p>
	<p>Further Input and modification:</p> <p>অদिति দৌড়ে গেলো তার মায়ের কাছে, "মা, দেখানো, বাবা আমাকে মেলায় নিয়ে যাচ্ছে না! কতো দেরি হয়ে গেলো বলো?" Aditi ran to her mother, "Mom, look, Dad isn't taking me to the fair! How much longer will it be?"</p>
Part 2	<p>Continuation:</p> <p>মা: (হেসে) বাবা তো একটু কাজ করছে, মা! তুই একটু ধৈর্য ধর। অদিতির বড় বোন, রিয়া: হ্যাঁ অদिति, একটু অপেক্ষা কর। আমরা সবাই মিলে যাবো মন্ডপে, আর তুই তো এত সুন্দর সেজেছিস! মা: ঠিক বলেছিস রিয়া। আর অদिति, মন্ডপে গিয়ে মা দুর্গার কাছে আমার জন্য একটু প্রার্থনা করিস তো, মা। অদिति: (উচ্ছ্বাসিত হয়ে) ঠিক আছে মা, আমি মা দুর্গার কাছে তোমার জন্য প্রার্থনা করবো! Mother: (Smiling) Dad is doing some work, dear. Be a little patient. Aditi's older sister, Riya: Yes, Aditi, wait a bit. We'll all go to the pandal together, and you look so beautiful! Mother: That's right, Riya. And Aditi, when you go to the pandal, please say a prayer to Goddess Durga for me. Aditi: (Excitedly) Okay, Mom, I'll pray to Goddess Durga for you!</p>
User Satisfaction Study : Case 2	
	<p>রাফি আর সন্দীপ ভালো বন্ধু। দুজন একসাথেই পাড়েছে অনার্স, এ বছরই পাশ করে বেরুবে। কিন্তু চাকরি খোঁজার বদলে তাদের মাথায় ভূত চেপেছে কফিশপ দেয়ার। Rafi and Sandeep are good friends. Both studied honors together and will graduate this year. But instead of looking for jobs, they've got the idea of starting a coffee shop stuck in their heads!</p>
Part 1	<p>Initialization:</p> <p>রাফি: সন্দীপ, চাকরি তো সবাই খোঁজে। আমরা একটু ভিন্ন কিছু করি, কফিশপ দিলে কেমন হয়? সন্দীপ: হ্যাঁ রাফি, আমাদের ইউনিক আইডিয়া যদি হয়, তাহলে সবাই আকৃষ্ট হবে। রাফি: ঠিক বলেছিস। ভাবছি, কফির সাথে কিছু স্পেশাল ডেসার্ট রাখব। তুই কী বলিস? সন্দীপ: দারুণ আইডিয়া! আর স্টুডেন্টদের জন্য ডিসকাউন্ট দিলে আরও ভালো চলবে। Rafi: Sandeep, everyone looks for jobs. How about we do something different and start a coffee shop? Sandeep: Yes, Rafi, if we have a unique idea, everyone will be attracted to it. Rafi: You're right. I'm thinking of offering some special desserts with the coffee. What do you think? Sandeep: Great idea! And if we offer discounts for students, it will do even better.</p>
	<p>Further Input and modification:</p> <p>সন্দীপের পিসেমশাই এর একটা বেকারি শপ আছে। তাই রাফির মনে হল উনার সাথে দেখা করে কিছু পরামর্শ নিলে ভালো হয়। যেই ভাবা সেই কাজ, পরদিন সন্ধ্যায় সন্দীপের পিসেমশাই-এর বাসায়। Sandeep's uncle has a bakery shop. So Rafi thought it would be good to meet him for some advice. As soon as they thought of it, they did it. The next evening, they were at Sandeep's uncle's house.</p>
Part 2	<p>Continuation:</p> <p>পিসেমশাই (আইডিয়া শ্রবণে): তোমাদের কফিশপের আইডিয়া শুনলাম। দারুণ তো। কোন ধরনের কফি রাখবে ভাবছো? রাফি: আমরা বিভিন্ন ধরনের কফি রাখতে চাই, সাথে কিছু স্পেশাল ডেসার্ট। আপনি যদি কিছু পরামর্শ দেন, খুব উপকার হবে। পিসেমশাই: অবশ্যই! আমাদের বেকারির কিছু রেসিপি তোমাদের কাজে লাগতে পারে। পিসিমা: আরেকটা কথা, গ্রাহকদের সাথে ভালো সম্পর্ক গড়া খুব জরুরি। তুমি খেয়াল রেখো। সন্দীপ: ঠিক বলেছে পিসিমা। আমরা সেটা মাথায় রাখব। Uncle (after hearing the idea): I heard about your coffee shop idea. It's great! What kind of coffee are you planning to offer? Rafi: We want to offer various types of coffee along with some special desserts. If you could give us some advice, it would be very helpful. Uncle: Of course! Some recipes from our bakery might be useful to you. Aunt: One more thing, building a good relationship with customers is very important. Make sure you pay attention to that. Sandeep: You're right, Aunt. We'll keep that in mind.</p>
User Satisfaction Study : Case 3	
	<p>আতিক ও আতিকা জমজ ভাই-বোন। পাড়েও একসাথে, ক্লাস ফোরো। গ্রীষ্মের ছুটি শুরু হয়েছে, একটু পরই মামাবাড়ি বেড়াতে নিয়ে যাবে তারা। অন্যদিন সারাদিন দুজনে মারামারি খুনসুটি করলেও আজ তারা দুজনেই চুপ, কেননা, বাবা বলেছে মারামারি করলে কোথাও বেড়াতে নিয়ে যাবে না। Atik and Atika are twin siblings. They study together in class four. Their summer vacation has started, and soon they will be visiting their uncle's house. Although they usually spend the whole day fighting and teasing each other, today they are both quiet because their father said he wouldn't take them anywhere if they fought.</p>
Part 1	<p>Initialization:</p> <p>আতিক: (ফিসফিসিয়ে) আতিকা, মামাবাড়ি গেলে রাহাত ভাইয়ার সাথে কত মজা করব, তাই না? আতিকা: হ্যাঁ, আর নতুন বানানো মিঠা-মিষ্টি খাবো। বাগানের আম তো আছেই! তুই কিছু আজ ঝগড়া করবি না। আতিক: (শুচকি হেসে) না, করব না। রাহাত ভাইয়া বলেছে এবার নদীতে গোসল করতে নিয়ে যাবে। Atik: (Whispering) Atika, it will be so much fun with Rahat bhaiya when we go to uncle's house, right? Atika: Yes, and we'll eat the sweets and cakes that grandma makes. And don't forget the mangoes from the garden! But you better not fight today. Atik: (Smiling) No, I won't. Rahat bhaiya said he would take us swimming in the river this time!</p>
	<p>Further Input and modification:</p> <p>নানাবাড়ি গিয়ে দেখে নানা বাইরে ঠাঁড়িয়ে ডাব কাটছে; আতিক ডাক দিলো, "নানা ভাইইইইইই!" When they arrived at their grandparents' house, they saw their grandfather outside cutting coconuts. Atik called out, "Grandpaaaaaaaa!"</p>
Part 2	<p>Continuation:</p> <p>নানা: হেসি মুখ? আরে, তোরা এলি? এতো দেরি কেনো? আতিকা, ডাব খাবে? আতিকা: (উচ্ছ্বাসিত) হ্যাঁ, নানা! নানী: (বাড়ির ভেতর থেকে) কে এসেছে, কে এসেছে? আরে আমার ছোট্ট আতিক আর আতিকা! তোমরা এসে গেছো? মামী: (খুশি হয়ে) এতক্ষণে এলে তোমরা! রাহাত: (আতিকা আর আতিককে দেখে) কি ভাই-বোন, আজ ঝগড়া হবে না তো? আতিক: (হাসতে হাসতে) না, রাহাত ভাই, আজ কোন ঝগড়া না। আগে তুমি বানো নদীতে কখন নিয়ে যাবে? আজ চলনো! Grandpa: (Smiling) Oh, you guys are here? Why are you so late? Atika, would you like a coconut? Atika: (Excitedly) Yes, Grandpa! Grandma: (From inside the house) Who's here, who's here? Oh, my little Atik and Atika! Have you both arrived? Aunt: (Happily) You finally made it! Rahat: (Seeing Atika and Atik) So, siblings, there won't be any fighting today, right? Atik: (Laughing) No, Rahat bhaiya, no fights today. But first, tell us when you'll take us to the river. Let's go today!</p>

Figure 2: All designed cases for the user study.

where correctness scores generally decrease and quality scores either drop or slightly increase in continuation parts. For instance, in Case 1, correctness falls from 3.51 (P1) to 3.40 (P2), and quality drops from 2.99 to 2.95. Case 2 shows a small decline in correctness from 3.48 to 3.45 but an increase in quality from 2.87 to 3.31. Similarly, in Case 3, correctness decreases from 3.34 to 3.30, while quality improves from 3.06 to 3.40. These patterns suggest that while the LLM’s correctness tends to diminish over continued interactions, the perceived quality of responses can vary, often improving or stabilizing in P2. This indicates a potential trade-off between maintaining accuracy and enhancing response quality in extended interactions.

বাবা: অদিতি, একটু ধৈর্য ধরো মা, আমি একটু কাজ শেষ করে নিচ্ছি। তারপর আমরা মতপে যাব।
 অদিতি: কিন্তু বাবা, আমি তো অনেকক্ষণ ধরে অপেক্ষা করছি!
 বাবা: আমি জানি মা, আর মাত্র দশ মিনিট। তুমি ততক্ষণে তোমার মা'কে গিয়ে দেখাও তো, তুমি কেমন সুন্দর করে সেজেছ।

C1P1-Correctness *

In this part, I thought the LLM worked well and correctly understood the dialect and related complexity.

1 2 3 4 5

Strongly Disagree Strongly Agree

Figure 3: A snapshot of the survey form (Case 1, Part 1).

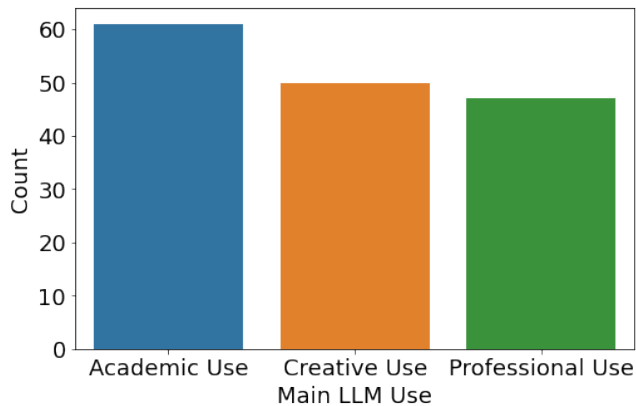


Figure 4: LLM use frequency of the participants.

4.3 Observation

LLMs show varying performance across different dialects, with noticeable improvements in quality in continuation parts for Cases 2 and 3. This suggests that the model may benefit from additional training on dialect-specific data to enhance its performance. The variability in standard deviations indicates inconsistent user experiences. Addressing these inconsistencies could improve overall satisfaction. Both Cases 2 and 3 show improved quality in continuation parts, suggesting that extending interactions can sometimes

enhance performance. However, the overall decrease in correctness highlights a need for better maintenance of accuracy in extended responses. To address the observed biases and performance issues, incorporating more diverse and representative dialect data could improve the LLM’s handling of various linguistic contexts.

5 Discussion

The user study results indicate a mixed performance of the LLM in handling creative storytelling across dialects. Overall satisfaction improved from a moderate pre-survey score of 3.31 to 3.92 post-interaction, demonstrating that users recognized some potential in the system. Correctness ratings were consistently higher than quality ratings, suggesting that while the LLM accurately processed input and maintained context, its narrative quality—especially in extended interactions—needs improvement. The continuation phases (P2) generally showed a decline in correctness compared to the initialization phases (P1), highlighting challenges in sustaining the initial performance level over time. Variability in user feedback, reflected by higher standard deviations in quality scores, points to inconsistent experiences across different cases. Notably, strong context awareness scores indicate that the system effectively retains dialogue context, a key strength in narrative generation. However, lower scores in dialect and cultural sensitivity underscore the necessity for more comprehensive dialect-specific training data. These findings collectively emphasize the need for balanced model tuning to enhance both factual accuracy and creative richness.

A case-wise analysis further reveals nuanced differences in performance based on dialectal settings. In Case 1, which focused on the West Bengal dialect, slight declines were observed in both correctness and quality during the continuation phase, suggesting stable yet modest performance. Conversely, Cases 2 (cross-dialect) and 3 (Bangladesh dialect) showed minor drops in correctness but notable improvements in quality over time, hinting at an adaptive narrative approach that might resonate better with users in these contexts. The variations in standard deviations across cases highlight that user experiences are not uniform, reinforcing the importance of addressing underlying inconsistencies. The data imply that extended interactions can either exacerbate or mitigate perceived quality, depending on the dialect context. This trade-off between maintaining high correctness and enhancing narrative quality remains a critical area for further refinement. Tailoring training strategies to account for specific cultural and linguistic nuances could help reduce biases and improve overall performance. These insights set the stage for targeted model enhancements in multilingual and culturally diverse applications.

5.1 Limitations

This study has several limitations. The user study’s sample size and demographic composition may not fully represent the diversity of native Bangla speakers. Our evaluation focused on specific creative storytelling scenarios, which might not capture the full spectrum of narrative challenges. Self-reported measures introduce potential biases that could affect the reliability of the feedback. The controlled environment of the study may not reflect real-world interactions with LLMs. Additionally, the exploration of dialectal variations was limited, leaving room for more comprehensive analysis. Future

research should address these gaps by incorporating larger, more diverse samples and broader evaluation contexts.

6 Concluding Remarks

In conclusion, this work underscores the complexities of deploying LLMs for creative storytelling in linguistically diverse environments like Bengali. The study demonstrates that while the LLM excels in maintaining context and delivering accurate initial responses, its performance tends to decline over extended interactions, particularly in terms of correctness. The observed gap between correctness and narrative quality suggests that creative storytelling requires not only factual accuracy but also sustained engagement and cultural sensitivity. Our case-wise analysis indicates that tailored approaches are necessary to address the unique challenges posed by different dialects, as evidenced by the varying performance across West Bengal, cross-dialect, and Bangladesh settings. Enhancing the model with more diverse dialectal data and refined feature engineering is crucial to mitigate biases and improve consistency. Overall, these findings contribute to a deeper understanding of the balance between creative freedom and accuracy in AI-generated content. Future research should focus on integrating culturally adaptive techniques to further enrich narrative quality. This work thus marks a significant step toward developing more inclusive and contextually aware LLMs for low-resource languages.

References

- [1] Anirudh Atmakuru, Jatin Nainani, Rohith Siddhartha Reddy Bheemreddy, Anirudh Lakkaraju, Zonghai Yao, Hamed Zamani, and Haw-Shiuan Chang. 2024. CS4: Measuring the Creativity of Large Language Models Automatically by Controlling the Number of Story-Writing Constraints. arXiv:2410.04197 [cs.CL] <https://arxiv.org/abs/2410.04197>
- [2] John Brooke. 1995. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* 189 (11 1995).
- [3] Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward Cultural Bias Evaluation Datasets: The Case of Bengali Gender, Religious, and National Identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*. 68–83.
- [4] Dipto Das, Carsten Østerlund, and Bryan Semaan. 2021. "Jol" or "Pani"?: How Does Governance Shape a Platform's Identity? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
- [5] Dipto Das and Bryan Semaan. 2022. Collaborative identity decolonization as reclaiming narrative agency: Identity work of Bengali communities on Quora. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.
- [6] Yuzhou Huang, Yiran Qin, Shunlin Lu, Xintao Wang, Rui Huang, Ying Shan, and Ruimao Zhang. 2024. Story3D-Agent: Exploring 3D Storytelling Visualization with Large Language Models. arXiv:2408.11801 [cs.CV] <https://arxiv.org/abs/2408.11801>
- [7] Mohammad Nurul Islam and Azirah Hashim. 2019. Historical evolution of English in Bangladesh. *Journal of Language Teaching and Research* 10, 2 (2019), 247–255.
- [8] Guillermo Marco, Luz Rello, and Julio Gonzalo. 2025. Small Language Models can Outperform Humans in Short Creative Writing: A Study Comparing SLMs with Humans and LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 6552–6570. <https://aclanthology.org/2025.coling-main.437/>
- [9] Yufei Tian, Tenghao Huang, Miri Liu, Derek Jiang, Alexander Spangher, Muhao Chen, Jonathan May, and Nanyun Peng. 2024. Are Large Language Models Capable of Generating Human-Level Narratives?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 17659–17681. <https://doi.org/10.18653/v1/2024.emnlp-main.978>
- [10] Azmine Toushik Wasi, Raima Islam, Mst Rafia Islam, Taki Hasan Rafi, and Dong-Kyu Chae. 2024. Exploring Bengali Religious Dialect Biases in Large Language Models with Evaluation Perspectives. In *The First Human-centered Evaluation and Auditing of Language Models Workshop, The ACM Conference on Human Factors in Computing Systems (CHI)*.
- [11] Azmine Toushik Wasi, Taki Hasan Rafi, and Dong-Kyu Chae. 2024. DiaFrame: A Framework for Understanding Bengali Dialects in Human-AI Collaborative Creative Writing Spaces. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose, Costa Rica) (CSCW Companion '24). Association for Computing Machinery, New York, NY, USA, 268–274. <https://doi.org/10.1145/3678884.3681862>