MRS Journal of Arts, Humanities and Literature
ISSN *(Online) 3049-1444*
Vol-3, Iss-11 (November-2026)
Journal homepage: https://mrspublisher.com/journals

Trans(formational) Challenges: Pragmatic and Cultural Loss in AI Translation of Bangla Texts into English

Fariah Hossain Labonno*

Lecturer, Department of English, Southeast University, Dhaka, Bangladesh

E-mail: fariah.hossainlabonno@seu.edu.bd

Received: 15 / 09 / 2026	Accepted: 24 / 10 / 2026	Published: 06 / 11 / 2026

Abstract: Large language models (LLMs) are now widely used to translate Bangla into English, and their output usually reads smoothly. This study asks whether that fluency preserves pragmatic and cultural meaning. It compares the translations of five LLM platforms (Gemini, ChatGPT, DeepSeek, Grok and Claude) with published human translations of 38 Bangla segments. Most segments come from three canonical texts: Jibanananda Das's "Banalata Sen", the opening chapter of Bibhutibhushan Bandyopadhyay's *Pather Panchali* and Sheikh Mujibur Rahman's 7 March speech. The remaining segments come from Shazed ul Hoq Khan Abir's short story "Mofiz-er Relation-e Spark Nai" (2026), a recent work that the models are unlikely to have encountered during "training." Drawing on implicature, speech act theory, House's model of register and Venuti's notion of domestication, the study reads each version closely against its source. The LLMs proved most reliable with names and famous texts, but least reliable with idioms, polysemous words and traditional counting units. In the speech, four of the five systems lost or reversed the force of a key threat. Changed numbers and substituted everyday objects were concentrated in the recent story, which suggests that strong results on canonical texts may partly reflect familiarity. The human translations were more idiomatic, yet they also domesticated the texts through added explanation and converted cultural units. The study therefore argues that human and machine translators both transform Bangla texts, though in different directions, and that fluent English is no guarantee that meaning has survived.

Keywords: *machine translation; large language models; Bangla; pragmatic equivalence; culture-specific items; domestication.*

Cite this article: Labonno, F. H. (2026). Trans(formational) Challenges: Pragmatic and Cultural Loss in AI Translation of Bangla Texts into English. *MRS Journal of Arts, Humanities and Literature, 3(11),1-8.*

Introduction

Not long ago, a Bangla speaker who needed a passage in English had to translate it or find someone who could. Today, the same task takes seconds. General-purpose large language models (LLMs) such as ChatGPT, Gemini and Claude translate on request, and their English usually reads well. For students, office workers and writers in Bangladesh, they have quietly become everyday translation tools.

Smooth English, however, is not the same as faithful English. The risk of a gap between the two is especially high for Bangla. Much Bangla meaning is left to be inferred from context, so a host who says *ekhane thanda* ("it's cold here") may in fact be asking for the window to be closed. Bangla also marks social relations that English leaves unmarked. Its three forms of "you", *apni*, *tumi* and *tui*, signal respect, familiarity or intimacy, whereas English has only one (Uddin, 2020). Idioms, kinship terms and local religious references add a further layer of cultural meaning that a word-for-word version cannot convey.

These are precisely the areas in which machine translation is weakest. Translation models tend to render idioms literally (Dankers et al., 2022), and LLMs remain unreliable with culture-specific items (Yao et al., 2024). Bangla, for its part, has long been underserved in language technology (Joshi et al., 2020), and current LLMs still perform worse on Bengali language and culture than on English (Kabir et al., 2025). Even so, most research on Bangla-English translation judges quality by automatic scores, which count shared words and reveal little about tone, politeness or implied meaning.

This study uses the term *trans(formational)* to describe what is at stake. When a text is translated, its meaning is transformed as it moves between languages. At the same time, the translation forms an image of that text, and of the culture behind it, for readers who meet it only in English. If an LLM turns a grafted orchard into "a garden of pens", the error does not stay on the page; it shapes how Bangla culture is understood abroad.

To examine this process, the study compares five LLM platforms with published human translations of Bangla poetry, prose and political speech. It also includes a recently published short story, which the models are unlikely to have seen, in order to test whether famous texts are translated better than unfamiliar ones. Throughout, the human translations are read as critically as the machine ones, since human translators also reshape the texts they carry across. The study addresses three research questions:

Copyright © 2026 The Author(s): This open-access article is distributed under the Creative Commons Attribution–NonCommercial 4.0 International License (CC BY-NC 4.0), permitting non-commercial use, sharing, and reproduction with proper credit to the author(s) and source.

1. To what extent do LLM translations of Bangla texts preserve pragmatic and cultural meaning, compared with published human translations?
2. Which kinds of meaning are most often lost or altered, and how does this vary across poetry, prose and speech?
3. Do LLMs translate canonical Bangla texts more successfully than a recently published one?

Literature Review

Meaning beyond words

Translation scholars have long held that a translator carries meanings rather than words. Nida (1964) placed the reader at the centre of this task, arguing that a good translation should affect its readers much as the original affected its own. House (2015) made the idea more precise. In her model, a translation succeeds when it keeps the function that the source text served in its setting, and that function is carried largely by register. Pragmatics explains what is lost when this function is missed. Grice (1975) showed that speakers often imply more than they say, while Austin (1962) and Searle (1976) showed that words also perform actions such as promising, declaring and threatening. Brown and Levinson (1987) added that speakers use politeness to protect one another's dignity, and that the forms this takes differ across cultures. Each of these layers can vanish even when every word is correctly translated. For this reason, Baker (2018) calls for *pragmatic equivalence*, by which a translation keeps not only what the source says but also what it implies and does.

Translation choices are also cultural and political. Venuti (1995) argued that translators working into English have often smoothed foreign texts into familiar patterns, so that the foreignness of the original becomes invisible. His warning applies in two directions here. LLM output may flatten Bangla texts into plain English, but published human translations, which are usually treated as the standard, may have been smoothed in the same way.

Features of Bangla that resist translation

Several features of Bangla carry social meaning that English grammar does not mark. The clearest is the three-way system of address noted above, which forces an English translator to signal respect or closeness through other means (Uddin, 2020). In Brown and Levinson's (1987) terms, such forms encode politeness in grammar itself, so their loss is not merely a matter of style. Bangla speakers also use kinship terms such as "brother" or "aunt" for neighbours and strangers, and these choices reflect class, power and region (Snigdha, 2022). In addition, Bangla has two written styles: the older literary *sadhu bhasha* and the standard *cholit bhasha*, alongside strong regional dialects. Ferguson (1959) described such splits between a formal and an everyday variety as *diglossia*. Two of the texts in this study show the contrast clearly, since *Pather Panchali* uses the older *sadhu* style, while the 7 March speech mixes standard Bangla with the speech of eastern Bengal.

Evaluating machine translation

Machine translation has become far more fluent, yet fluency has proved a poor guide to quality. The move to neural models in the mid-2010s sharply improved readability (Castilho et al., 2017), and LLMs now translate well between widely used languages, although less well where data is scarce (Hendy et al., 2023). Claims that machines have reached human quality, however, have not survived careful testing. Judges prefer human translation more strongly when they read whole documents rather than isolated sentences (Läubli et al., 2018), and verdicts also shift with the judges' expertise and the test material (Toral et al., 2018). For this reason, Läubli et al. (2020) recommend that translations be judged in context, using whole documents and original rather than translated source texts. Automatic metrics such as BLEU (Papineni et al., 2002) add a further problem, because they measure word overlap with a reference translation. Freitag et al. (2021) showed that such scores can rank systems differently from expert judges.

A deeper question is whether these systems grasp meaning at all. Bender and Koller (2020) argue that a system trained only on text never connects language to the world or to a speaker's intention. One need not accept this argument fully to see what it predicts: errors should gather where meaning depends on context and shared knowledge rather than on textual patterns.

Where machine translation fails

Research confirms that the weak points of machine translation lie mostly above the level of the word. Dankers et al. (2022) found that translation models process idioms word by word, producing literal versions of figurative phrases. Troiano et al. (2020) found that the emotions of a source text are often lost or altered in machine translation. Similarly, Yao et al. (2024) showed that LLMs remain weak on Asian cultural concepts and tend to invent information when a term is rare. Prates et al. (2020) found that Google Translate turned the gender-neutral pronouns of twelve languages, Bangla among them, into "he" far more often than "she". Such errors matter beyond literature, since users of machine translation in hospitals and courts often know too little about its limits (Vieira et al., 2021).

Literary texts pose a particular challenge. Toral and Way (2018) found that readers judged only 17 to 34 per cent of machine-translated sentences from English novels to be as good as a professional translation. More recently, Karpinska and Iyyer (2023) showed that an LLM translated literary passages better when given whole paragraphs, although serious errors such as omissions persisted. Neither study, however, included Bangla, poetry or political speech.

Human translators and LLMs compared

Studies that compare human and machine translators on culturally rich material reach a consistent verdict: machines convey the basic content but lose the figurative and cultural layer. Kolesnikova (2025), for example, found that ChatGPT, DeepL and Google Translate copied Russian idioms word for word and lost their imagery. Megdadi et al. (2026) compared professional translators, ChatGPT and Gemini on 40 English-Arabic food idioms; the human translators scored highest on every measure, and the widest gap lay in cultural adaptation. Likewise, Sukaesih et al. (2026) found that ChatGPT translated Sundanese cultural terms largely accurately but rarely explained them, whereas a human translator often added a short gloss. For Bengali, Singh et al. (2025) reported that LLMs prompted simply to translate idioms received the lowest ratings from human judges.

These studies nevertheless share two limits. Most test isolated items rather than whole texts, in which tone and voice also matter. More importantly, most treat the human translation as the correct answer, so they cannot detect cases in which a machine stays closer to the original.

Bangla in language technology

Bangla is among the most widely spoken languages in the world (Eberhard et al., 2023), yet Joshi et al. (2020) placed it among languages with a large online presence but little carefully prepared data. Successive work narrowed this gap, from statistical and neural systems (Mumin et al., 2019a, 2019b) to a large parallel corpus (T. Hasan et al., 2020), back-translated training data (Mondal et al., 2024) and a single model covering 200 languages (NLLB Team, 2024). Each step raised automatic scores, but none asked whether translations kept politeness, implication or cultural reference.

Recent studies suggest that these are exactly the weak areas. Sijan et al. (2026) found that LLMs often chose overly formal address and confused kinship terms in Bangladeshi social interaction, and Shuvo et al. (2026) found that models mixed *apni* and *tumi* within a single formal text. In a related study, M. A. Hasan et al. (2026) showed that models asked about Bangla culture in English tend to replace local references with dominant global ones. These findings reflect a wider imbalance. The quality of language technology follows economic power more than need (Blasi et al., 2022), and systems built around dominant languages struggle to express ideas from other cultures (Helm et al., 2023). Bird (2020) therefore warns against treating a community's language simply as data to be extracted.

Research gap

Three gaps follow from this review. First, Bangla-English translation research has relied on automatic scores, while recent Bangla studies of LLMs test cultural knowledge through question answering rather than translation (Kabir et al., 2025; Sijan et al., 2026). Second, comparative studies mostly test isolated items rather than literary and spoken texts. Third, they usually treat the human translation as correct. To the author's knowledge, no study has yet used close reading to compare current LLMs with published human translations of Bangla, examining both with equal scrutiny. The present study addresses these gaps.

Theoretical Framework

The analysis draws on four connected perspectives, each of which asks a different question of a translation. Taken together, they turn the idea of *trans(formational)* translation into questions that can be put to every segment.

The first perspective concerns what a text implies and does. According to Grice (1975), conversation rests on a cooperative principle, and speakers who openly break its maxims signal an implied meaning. A poet who claims to have walked for a thousand years is not lying; rather, the reader infers weariness and the weight of time. Speech act theory adds that utterances perform actions, and that their force can differ from their literal form (Austin, 1962; Searle, 1976). A translation that keeps the content of a threat but weakens its force has therefore changed what the text does.

The second perspective concerns register. House (2015) judges a translation by comparing its *field* (subject matter), *tenor* (the relationship between speaker and audience) and *mode* (the channel, spoken or written) with those of the source. She also distinguishes *overt* translations, which remain openly foreign, from *covert* ones, which read as if first written in the target language. Historical and literary texts such as those studied here call for overt translation, because readers value them for their origin. A version of the 7 March speech that reads like a modern press release has thus failed in tenor and mode, however accurate its content.

The third perspective concerns form and effect. Nida (1964) distinguished *formal equivalence*, which keeps the form of the source, from *dynamic equivalence*, which seeks a similar effect on the reader. Literary translators often favour the latter, whereas LLMs asked for a translation tend toward the former. A comparison may therefore reward a human version simply for being freer, so this study keeps closeness to the source and preservation of effect as separate questions.

The fourth perspective concerns direction. Venuti (1995) contrasted *domesticating* translation, which brings the text toward the reader, with *foreignising* translation, which brings the reader toward the text. The analysis applies this lens to both human and machine versions and records the direction of each shift without assuming that either strategy is wrong. Finally, research on language inequality (Blasi et al., 2022; Helm et al., 2023) helps to explain why errors fall where they do. It predicts that errors should gather in idioms, regional speech and culture-specific terms, since these are the features of Bangla least represented in training data.

Methodology

Design and texts

This study uses a qualitative comparative design. Each Bangla segment is read alongside six English versions, one published human translation and five LLM translations, in order to establish what each keeps, changes or loses. Close comparison of source and target is a long-established method for explaining how meaning shifts in translation (House, 2015).

The corpus contains 38 segments in two tiers. Tier 1 contains 24 segments, eight from each of three canonical texts. The first is Jibanananda Das's poem "Banalata Sen", in Fakrul Alam's translation (Das, 1999). The second is the opening chapter of Bibhutibhushan Bandyopadhyay's novel *Pather Panchali*, in the translation by T. W. Clark and Tarapada Mukherji (Bandyopadhyay, 1929/1968). The third is Sheikh Mujibur Rahman's speech of 7 March 1971. Its Bangla text was taken from *Prothom Alo* (2021), and its English version is the widely circulated one credited to the Department of Films and Publications (n.d.). Tier 2 contains 14 segments from two paragraphs of Shazed ul Hoq Khan Abir's short story "Mofiz-er Relation-e Spark Nai" and its English translation by Arunava Sinha and Shabnam Nadiya (Abir, 2026a, 2026b). Since both versions were published on *adda* in August 2026, the story is very unlikely to appear in the models' training data, whereas the canonical texts almost certainly do. Comparing the two tiers therefore offers a first indication of whether familiarity affects translation quality.

Each segment is one to three sentences of prose or one to four lines of verse. Every segment contains at least one feature known to be difficult to translate, such as an idiom, a form of address, a culture-specific item, a figurative image or a rhetorical device. All segments and versions are provided as supplementary material.

Data collection

Five widely used platforms were tested: Gemini (Google), ChatGPT (OpenAI), DeepSeek (DeepSeek), Grok (xAI) and Claude (Anthropic). To reflect ordinary use, each was accessed through its free public interface with the default model at the time:

Gemini 3.5 Flash-Lite, GPT-6 Luna, DeepSeek V4, Grok 4 Fast and Claude Sonnet 5.5. Following Läubli et al.'s (2020) advice to evaluate translation in context, each text was entered as a connected passage in a new session with default settings, and the segments were then taken from the resulting translations. The same prompt was used throughout, with no information about author or genre: "Translate the following Bangla text into natural English, preserving tone and cultural meaning." Because the prompt explicitly requested tone and cultural meaning, any loss of these cannot be attributed to the instruction. All outputs were collected between 28 and 30 October 2026. In an earlier trial, the Tier 2 sentences were also entered one at a time; this run is not part of the main data, but one contrast from it is reported because it shows the effect of context.

Analysis

The analysis proceeded in three steps. First, the difficult feature of each segment was identified from the Bangla alone, before any translation was read, so that later judgements were anchored in the source rather than in how smoothly a version read. Second, all six versions were read against the Bangla using the four questions set out in the Theoretical Framework, and each was recorded as keeping, weakening or losing the feature, with any added content noted separately. Third, these records were compared across translators, genres and tiers to identify recurring patterns. Segments in which the versions diverged sharply, and those in which most LLMs failed in the same way, were selected for discussion, and every segment discussed is quoted so that the readings can be checked.

Findings

Across all four texts, the five LLMs produced fluent English. Their difficulties lay instead in idioms, culture-specific words, counting systems and the force of spoken language. The published human translations also changed the texts, although in different ways.

Poetry: close agreement and quiet losses

The six versions of "Banalata Sen" were strikingly similar. Most kept the place names, the extended sea simile and the closing image, and DeepSeek's opening line, "For a thousand years I have walked the ways of the world", matched Alam's published translation word for word. Such closeness to a well-known translation suggests that the poem is familiar to the models.

The losses that did occur were small but revealing. In "আমারে দু-দণ্ড শান্তি দিয়েছিলো", the speaker receives peace for only *du-donda*, a brief measured span. All five LLMs kept this limit as "a moment's peace", whereas Alam's "Finally found some bliss" removes it. Likewise, Alam renders *sondhya* (evening) as "Night falls", while every LLM keeps "evening". In both places, the human translator chose a more musical line at the cost of a small shift in meaning. By contrast, one loss affected all six versions. When Banalata Sen asks "এতদিন কোথায় ছিলেন?", she uses the respectful verb form *chhilen*, and this respect is part of the poem's tenderness. No version could carry it, because English "you" has only one form; the loss is structural rather than individual. The most serious machine error in the poem was Claude's omission of Vidisha from "বিদিশার নিশা", which keeps the mood of the line but removes the ancient city that gives the image its historical depth.

Prose: culture-specific items under strain

Pather Panchali produced the widest variation, largely because its *sadhu*-style opening is dense with rural and social terms. The clearest case is "নতুন কলমের বাগান", a new orchard of grafted trees, in which *kolom* means "graft" but also "pen":

> Gemini: "a new grafted garden"
>
> Grok: "a new orchard"
>
> ChatGPT: "a new grove of young saplings"
>
> DeepSeek: "a new orchard of hybrid plants"
>
> Claude: "a new garden of pens"

Only Gemini preserved the practice of grafting. While Grok and ChatGPT produced a plausible orchard, DeepSeek introduced a modern "hybrid", and Claude chose the everyday sense of the word and produced nonsense. The human version, "some new mango trees", also departs from the text, since it names a fruit the Bangla does not mention and adds that the trees "died".

A related problem arose when common nouns were read as names. Where the Ichamati sweeps away the indigo planters "like straw" (*kutar moto*), Grok and Claude created a person called "Kutar", DeepSeek wrote "a hut", and ChatGPT omitted the word. ChatGPT similarly turned *nal phul* (water lilies) into "the Nal family". Such errors are difficult for an English reader to detect, because the result still reads as a sentence about people and places. Social terms fared little better. Only Gemini and Claude kept *kulin*, the high-ranking Brahmin whose custom of multiple marriages drives the passage, whereas the other three reduced him to "a renowned nobleman", erasing the practice the novel criticises.

The human translation handled these terms in a different way. Clark and Mukherji added a full explanatory sentence, "Kulins were notorious for multiple marriages", and converted the Bengali year 1240 into "1833", while all five LLMs kept "1240". They also rendered "ভাল মনে করিতেই পারে না" (could hardly think well of him) as "had very little memory of her husband", whereas every LLM kept the sense of resentment. On these points the machines stayed closer to the source. Nevertheless, the human version was clearly stronger in narrative voice. It keeps the habitual sense of the old woman's daily gesture ("she used to raise her hand ... she would say"), while all five LLMs turned this repeated scene into a single moment in the simple present. ChatGPT also added details absent from the text, giving her "a trembling hand" and calling her a "widow".

Speech: force, idiom and voice

The 7 March speech tested whether translations could preserve action as well as content. The most revealing segment is the threat "আমরা ভাতে মারব, আমরা পানিতে মারব", in which the idiom *bhate mara* means to kill by cutting off food. The line thus warns that the people will besiege the army by denying it food and water:

> Human: "We will starve them to death. We will make them go without water and choke them to death."
>
> Gemini: "We will starve them of food, we will drown them in water."
>
> ChatGPT: "We will fight with rice, we will fight with water."

DeepSeek: "We will strike with rice, we will strike with water."

Grok: "We will strike on land, we will strike on water."

Claude: "We will die in hunger, we will die in water."

Only the human version and Gemini recover the idiom, and Gemini loses it again by turning siege into drowning. ChatGPT and DeepSeek translate the words but not the meaning, while Grok substitutes a military cliché. Most strikingly, Claude reverses the speech act, so that a threat against the enemy becomes a statement that the speakers themselves will die. In Searle's (1976) terms, the illocutionary force is lost or inverted in four of the five machine versions. The human version keeps the force, although it adds "choke them to death", which the Bangla does not contain.

The speaker's voice, on the other hand, fared better with the machines. Four LLMs kept the emphatic hesitation in "আমি, আমি প্রধানমন্ত্রিত্ব চাই না" as "I, I do not want ...", whereas the human version and ChatGPT dropped it, turning live speech into written statement. The human version also altered the tenor of the address to soldiers: "তোমরা আমার ভাই" (you are my brothers) became "You are our brothers", and the personal "at my chest" became "shoot us". In addition, it domesticated the setting, rendering *gorur gari* (bullock carts) as "hackney carriages", a term from British urban transport, while all five LLMs kept "bullock carts" or "ox-carts". Domestication was not limited to the human translator, however, since ChatGPT and Claude converted *sat koti* (seven crore) into "seventy million". Finally, no version kept the regional verb forms of the closing lines, such as *dabaye rakhte parba na*, so every English reader meets a more standard speaker than the one heard in 1971.

The recent story: errors of basic meaning

In Abir's story, which the models are unlikely to have seen, LLM errors extended from idioms to ordinary words and numbers. The traditional counting unit *kuri* (twenty) is the clearest example. For *tin kuri* (sixty), Claude wrote "thirty", and for *panch kuri* (a hundred), ChatGPT wrote "sixty" and Claude "about fifty". The other systems gave both numbers correctly, with DeepSeek using the archaic "three score" and "five score". Notably, each wrong number appeared in a smooth, credible sentence.

Further lexical errors followed the same pattern. Claude rendered *dab* (green coconut) as "jackfruits", *kodal* (spade) as "hoe" and *thikuji-kuluji* (horoscope and family line) as "distant relatives", while Grok translated *binimagna* (unpaid) as "without even being asked". Idioms again produced literal versions. The phrase *tinkule keu nei* means that a person has no family at all. Gemini and ChatGPT conveyed this, but Grok, DeepSeek and Claude wrote "no one in three generations" or "in three lineages", which an English reader cannot interpret. By contrast, all five systems kept the folk deities Gazi Pir, Manasa and Bonbibi by name.

Context made a clear difference. In the earlier sentence-by-sentence trial, four of the five systems rendered *sundari*, the mangrove tree that gives the Sundarbans its name, as "beautiful trees". When the same sentence was translated within its paragraph, all five gave "Sundari trees". The human translation of the story was the most accurate of the six, yet it too reshaped the text. For instance, it rendered "under the NGO's own management" as "at the instigation of the NGO", which implies a hostility that the Bangla does not state.

Discussion

Fluency conceals loss

The first research question asked how far LLM translations preserve pragmatic and cultural meaning. The findings show that they do so unevenly and that their failures are hard to see. Every machine version reads as competent English, yet "a garden of pens", "we will die in hunger" and "thirty other labourers" are not matters of style; they change what the text says or does. This supports earlier evidence that fluency is a poor guide to translation quality (Castilho et al., 2017; Läubli et al., 2018), and it adds a specific risk for less-resourced languages. A reader with no Bangla receives no signal that a threat has been reversed or a number changed.

The systems also differed considerably. Gemini most often preserved culture-specific meaning, while Claude produced the largest number of serious errors, including the only reversal of a speech act. Such variation means that general claims about "AI translation" of Bangla should be treated with caution, since the choice of platform alone can decide whether a reader receives the meaning or its opposite.

The location of loss

The second research question concerned which kinds of meaning are lost and how this varies by genre. The pattern matches what earlier research predicts. Idioms were translated word by word, as Dankers et al. (2022) observed, and culture-specific items were replaced with general or invented ones, as Yao et al. (2024) found for Asian cultural concepts. Errors also gathered where meaning depends on knowledge beyond the text, such as a word that means both "graft" and "pen" or a counting unit used mainly in rural speech. This is consistent with Bender and Koller's (2020) argument. The genres, however, failed differently. The poem lost little beyond a structural honorific, the novel lost mainly cultural content, and the speech lost mainly pragmatic force. The last of these echoes Troiano et al.'s (2020) finding that machine translation often fails to carry the emotional charge of a source. In House's (2015) terms, the novel suffered most in field, while the speech suffered most in tenor and mode.

Familiar and unfamiliar texts

The third research question asked whether canonical texts are translated more successfully than a recent one. The evidence points in this direction, although it cannot settle the question. The versions of "Banalata Sen" were close to one another and, in one line, identical to Alam's translation, whereas the recent story produced changed numbers and substituted objects that did not appear in the poem. This contrast is consistent with Chang et al. (2023), who found that LLMs memorise popular books and perform better on them in downstream tasks. If so, studies that test LLMs only on canonical literature may overestimate how well they handle the everyday and regional Bangla that most users actually bring to them. The effect of context points the same way: the models did better with whole paragraphs than with single sentences, which agrees with Karpinska and Iyyer (2023).

Rethinking the human benchmark

The findings also question the common practice of treating the human translation as the correct answer (e.g. Megdadi et al., 2026). The published translations were the most idiomatic and the strongest in narrative voice, yet they took liberties that the

machines did not. They added explanation, converted the calendar year, traded a precise time limit for a smoother line, and replaced bullock carts with hackney carriages. These choices fit Venuti's (1995) account of domestication. The LLMs, prompted to preserve cultural meaning, more often did the opposite by keeping "1240", "Kulin" and "bullock carts". In this narrow sense, the machines were more foreignising than the human translators, a result that echoes Sukaesih et al.'s (2026) finding that ChatGPT carried cultural terms across unexplained. The difference, then, is not that one translator is faithful and the other free. Instead, the human translators shaped Bangla to read as good English literature, while the machines kept its surface and lost much of its depth. This double movement is what the term *trans(formational)* is meant to capture.

The concentration of errors in rural vocabulary, counting systems and idioms also reflects wider language inequality. Language technology performs best for the varieties that dominate its training data (Blasi et al., 2022; Joshi et al., 2020), and formal Bangla and famous literature are far better represented online than the village speech of the Sundarbans. When systems meet such material, they tend to replace it with something more familiar (Helm et al., 2023). "Jackfruits" for green coconuts and "beautiful trees" for the *sundari* mangrove are small instances of this process, but repeated across many texts and readers, such substitutions shape how Bangla culture appears in English.

Implications and limitations

These findings have practical value. Teachers of English and translation in Bangladesh can use them to show why a fluent machine translation still needs a bilingual reader, and which features to check first: idioms, numbers, polysemous words and forms of address. Users can draw a simple lesson, namely that LLMs are safer with standard Bangla than with regional texts and that paragraph input works better than single sentences. Developers, for their part, can see specific gaps, such as traditional counting units, that automatic scores are unlikely to reveal.

The study also has limitations. The analysis was conducted by a single bilingual analyst, so the readings, although quoted in full and open to checking, have not been tested for agreement among independent readers. The sample is small, and Tier 2 comes from a single story, which makes the comparison between tiers suggestive rather than conclusive. Each platform was tested once, and LLM outputs vary between sessions and change as models are updated. Finally, Claude, one of the systems tested, also assisted in preparing this article. Its translations were collected and reported under the same conditions as the others, but readers should be aware of this overlap.

Conclusion

This study compared five LLM platforms with published human translations of Bangla poetry, prose, political speech and a recent short story. The machines produced fluent English throughout, but fluency concealed real loss. Idioms were translated word for word, culture-specific terms were misread and numbers were changed. In one case, a threat against an enemy became a statement of the speakers' own death. Changed numbers and substituted everyday objects were concentrated in the recent story, which suggests that strong performance on famous texts may partly reflect familiarity rather than understanding. At the same time, the published human translations were not neutral. Their translators explained, converted and smoothed the texts toward English norms. Translation, in other words, always transforms meaning and forms an image of the culture behind it; what differs between human and machine translators is the direction and kind of transformation. Future research should test more texts and genres, compare several human translations of each source, and involve multiple independent bilingual readers. Until then, a smooth English version of a Bangla text should not be taken as evidence that its meaning has survived.

Funding

This research received no external funding. It was self-funded by the author.

Declaration of interest

The author declares no conflict of interest.

Use of generative AI

Generative AI assistance was used for limited language refinement and editorial purposes to improve sentence clarity and readability. All revisions were reviewed and verified by the author, who takes full responsibility for the accuracy, originality, and final content of the manuscript.

Data availability

All Bangla segments and the six English versions of each are available from the author on request.

Ethics

The study involved no human participants and did not require ethical approval.

References

1. Abir, S. u. H. K. (2026a, August 17). মফিজের রিলেশনে স্পার্ক নাই [No spark in Mafiz's relationship]. *adda*. Commonwealth Foundation. https://www.addastories.org/%e0%a6%ae%e0%a6%ab%e0%a6%bf%e0%a6%9c%e0%a7%87%e0%a6%b0-%e0%a6%b0%e0%a6%bf%e0%a6%b2%e0%a7%87%e0%a6%b6%e0%a6%a8%e0%a7%87-%e0%a6%b8%e0%a7%8d%e0%a6%aa%e0%a6%be%e0%a6%b0%e0%a7%8d%e0%a6%95-%e0%a6%a8/
2. Abir, S. u. H. K. (2026b, August 17). No spark in Mafiz's relationship (A. Sinha & S. Nadiya, Trans.). *adda*. Commonwealth Foundation. https://www.addastories.org/no-spark-in-mafizs-relationship/
3. Austin, J. L. (1962). *How to do things with words*. Clarendon Press.
4. Baker, M. (2018). *In other words: A coursebook on translation* (3rd ed.). Routledge.
5. Bandyopadhyay, B. (1968). *Pather Panchali: Song of the road* (T. W. Clark & T. Mukherji, Trans.). George Allen & Unwin. (Original work published 1929)
6. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). ACL. https://doi.org/10.18653/v1/2020.acl-main.463

7. Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 3504–3519). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.313

8. Blasi, D., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 5486–5505). ACL. https://doi.org/10.18653/v1/2022.acl-long.376

9. Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.

10. Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, *108*, 109–120. https://doi.org/10.1515/pralin-2017-0013

11. Chang, K. K., Cramer, M., Soni, S., & Bamman, D. (2023). Speak, memory: An archaeology of books known to ChatGPT/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7312–7327). ACL. https://doi.org/10.18653/v1/2023.emnlp-main.453

12. Dankers, V., Lucas, C., & Titov, I. (2022). Can transformer be too compositional? Analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 3608–3626). ACL. https://doi.org/10.18653/v1/2022.acl-long.252

13. Das, J. (1999). *Jibanananda Das: Selected poems with an introduction, chronology and glossary* (F. Alam, Trans.). The University Press Limited.

14. Department of Films and Publications. (n.d.). *The speech text* [English translation of the speech of 7 March 1971]. 7th March Foundation. Retrieved November 4, 2026, from https://7thmarch.com/the-speech-text/

15. Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2023). *Ethnologue: Languages of the world* (26th ed.). SIL International.

16. Ferguson, C. A. (1959). Diglossia. *Word*, *15*(2), 325–340. https://doi.org/10.1080/00437956.1959.11659702

17. Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (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. https://doi.org/10.1162/tacl_a_00437

18. Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Vol. 3. Speech acts* (pp. 41–58). Academic Press.

19. Hasan, M. A., Naswan, R., Samir, F., Sultana, S., & Ahmed, S. I. (2026). *When English rewrites local knowledge: Global narrative dominance in large language models* (arXiv:2605.30481). arXiv. https://arxiv.org/abs/2605.30481

20. Hasan, T., Bhattacharjee, A., Samin, K., Hasan, M., Basak, M., Rahman, M. S., & Shahriyar, R. (2020). Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 2612–2623). ACL. https://doi.org/10.18653/v1/2020.emnlp-main.207

21. Helm, P., Bella, G., Koch, G., & Giunchiglia, F. (2023). *Diversity and language technology: How techno-linguistic bias can cause epistemic injustice* (arXiv:2307.13714). arXiv. https://doi.org/10.48550/arXiv.2307.13714

22. Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). *How good are GPT models at machine translation? A comprehensive evaluation* (arXiv:2302.09210). arXiv. https://doi.org/10.48550/arXiv.2302.09210

23. House, J. (2015). *Translation quality assessment: Past and present*. Routledge.

24. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). ACL. https://doi.org/10.18653/v1/2020.acl-main.560

25. Kabir, D., Mahim, M. R. C., Shafayat, S., Sadik, A., Ahmed, A., Kim, E., & Oh, A. (2025). BLUCK: A benchmark dataset for Bengali linguistic understanding and cultural knowledge. In *Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025)* (pp. 161–179). ACL. https://aclanthology.org/2025.banglalp-1.14/

26. Karpinska, M., & Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 419–451). ACL. https://doi.org/10.18653/v1/2023.wmt-1.41

27. Kolesnikova, M. P. (2025). Interpretation of idioms and culturally marked elements in neural network translation (based on Russian-English parallels). *Litera*, (11), 187–216. https://doi.org/10.25136/2409-8698.2025.11.76823

28. Läubli, S., Castilho, S., Neubig, G., Sennrich, R., Shen, Q., & Toral, A. (2020). A set of recommendations for assessing human–machine parity in language translation. *Journal of Artificial Intelligence Research*, *67*, 653–672. https://doi.org/10.1613/jair.1.11371

29. Läubli, S., Sennrich, R., & Volk, M. (2018). Has machine translation achieved human parity? A case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791–4796). ACL. https://doi.org/10.18653/v1/D18-1512

30. Megdadi, S., Almahasees, Z., Al-Harahsheh, A., Al Rousan, R., Al-Taher, M. A., & Husienat, I. (2026). Translating food idioms across the English-Arabic language divide: A comparative study of human translators, ChatGPT, and Gemini. *Research Journal in Advanced Humanities*, *7*(2). https://doi.org/10.58256/pf7v7z49

31. Mondal, S. K., Wang, C., Chen, Y., Cheng, Y., Huang, Y., Dai, H.-N., & Kabir, H. M. D. (2024). Enhancement of English-Bengali machine translation leveraging back-translation. *Applied Sciences, 14*(15), 6848. https://doi.org/10.3390/app14156848

32. Mumin, M. A. A., Seddiqui, M. H., Iqbal, M. Z., & Islam, M. J. (2019a). shu-torjoma: An English↔Bangla statistical machine translation system. *Journal of Computer Science, 15*(7), 1022–1039. https://doi.org/10.3844/jcssp.2019.1022.1039

33. Mumin, M. A. A., Seddiqui, M. H., Iqbal, M. Z., & Islam, M. J. (2019b). Neural machine translation for low-resource English-Bangla. *Journal of Computer Science, 15*(11), 1627–1637. https://doi.org/10.3844/jcssp.2019.1627.1637

34. Nida, E. A. (1964). *Toward a science of translating*. E. J. Brill.

35. NLLB Team. (2024). Scaling neural machine translation to 200 languages. *Nature, 630*, 841–846. https://doi.org/10.1038/s41586-024-07335-x

36. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). ACL. https://doi.org/10.3115/1073083.1073135

37. Prates, M. O. R., Avelar, P. H., & Lamb, L. C. (2020). Assessing gender bias in machine translation: A case study with Google Translate. *Neural Computing and Applications, 32*(10), 6363–6381. https://doi.org/10.1007/s00521-019-04144-6

38. Prothom Alo. (2021, December 17). ৭ মার্চের ঐতিহাসিক ভাষণ [The historic 7 March speech]. Retrieved November 4, 2026, from https://www.prothomalo.com/bangladesh/৭-মার্চের-ঐতিহাসিক-ভাষণ-2

39. Searle, J. R. (1976). A classification of illocutionary acts. *Language in Society, 5*(1), 1–23. https://doi.org/10.1017/S0047404500006837

40. Shuvo, M. A., Hasan, M., Parvez, M. T., Noman, A. H., & Ovi, M. S. H. (2026). *Polite on the surface, wrong in practice: A curated dataset for fixing honorific failures in multilingual Bangla generation* (arXiv:2605.22487). arXiv. https://arxiv.org/abs/2605.22487

41. Sijan, T. A., Rifat, S. M. G., Partha, P. C., Islam, M. T., & Anwar, M. M. (2026). BanglaSocialBench: A benchmark for evaluating sociopragmatic and cultural alignment of LLMs in Bangladeshi social interaction. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Student Research Workshop)* (pp. 247–280). ACL. https://arxiv.org/abs/2603.15949

42. Singh, P. R., Prasad, K., Zaki, M., & Wasnik, P. (2025). *Graph-assisted culturally adaptable idiomatic translation for Indic languages* (arXiv:2505.21937). arXiv. https://arxiv.org/abs/2505.21937

43. Snigdha, S. S. (2022). Representation of social class and hierarchy in Bangla address terms: A sociolinguistic study. *Journal of Modern Languages, 32*(2), 107–129. https://doi.org/10.22452/jml.vol32no2.6

44. Sukaesih, I., Rakhmi, F. P., Humolungo, F., Indrayani, S., & Abdillah, T. E. (2026). A comparative evaluation of ChatGPT and human translations of Sundanese cultural terms. *Studies in English Language and Education, 13*(2), 1100–1115. https://doi.org/10.24815/siele.v13i2.1108

45. Toral, A., & Way, A. (2018). What level of quality can neural machine translation attain on literary text? In J. Moorkens, S. Castilho, F. Gaspari, & S. Doherty (Eds.), *Translation quality assessment: From principles to practice* (pp. 263–287). Springer. https://doi.org/10.1007/978-3-319-91241-7_12

46. Toral, A., Castilho, S., Hu, K., & Way, A. (2018). Attaining the unattainable? Reassessing claims of human parity in neural machine translation. In *Proceedings of the Third Conference on Machine Translation* (pp. 113–123). ACL. https://doi.org/10.18653/v1/W18-6312

47. Troiano, E., Klinger, R., & Padó, S. (2020). Lost in back-translation: Emotion preservation in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 4340–4354). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.384

48. Uddin, M. A. (2020). Second person pronouns as person deixis in Bengali and English: Linguistic forms and pragmatic functions. *International Journal of English Linguistics, 10*(1), 345–351. https://doi.org/10.5539/ijel.v10n1p345

49. Venuti, L. (1995). *The translator's invisibility: A history of translation*. Routledge.

50. Vieira, L. N., O'Hagan, M., & O'Sullivan, C. (2021). Understanding the societal impacts of machine translation: A critical review of the literature on medical and legal use cases. *Information, Communication & Society, 24*(11), 1515–1532. https://doi.org/10.1080/1369118X.2020.1776370

51. Yao, B., Jiang, M., Bobinac, T., Yang, D., & Hu, J. (2024). Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 13078–13096). ACL. https://aclanthology.org/2024.findings-emnlp.765