A MODEL FOR ASSESSING THE LINGUOCULTURAL ADEQUACY OF AI-GENERATED FOLKLORE TEXTS

Fayyozjon Azimov
Shahrisabz State Pedagogical Institute
fayyozjonazimov@gmail.com
Tel: +99891 811 41 77

Abstract

Large language models (LLMs) such as ChatGPT, Claude and Gemini now retell, translate and invent folk tales in seconds, and their texts are already used in classrooms, children's books and online media. Yet there is no established way to judge whether such a text is culturally adequate. Standard measures of machine translation quality, such as BLEU or general fluency ratings, can tell us whether a text is grammatical and close to a reference, but not whether the Dev is still a Dev, whether Zumrad's *sabr* is still a virtue, or whether the hero still greets the monster as *bobo*. Research on AI-generated folklore has shown that cultural losses occur at many levels, from names and idioms to metaphors, values, emotions, speech etiquette, narrative structure and visual imagery, but these findings have not been integrated into a practical evaluation tool.

This article proposes and validates the **Linguocultural Adequacy Assessment Model (LCAAM)** for AI-generated folklore texts. The model was developed in three phases. First, dimensions and indicators were derived from linguocultural, cognitive, pragmatic and narratological theory and from a series of empirical studies of English and Uzbek fairy tales retold by AI. Second, the model was refined and validated by a panel of experts through a two-round Delphi procedure, using content validity indices. Third, the model was piloted on a sample of AI-generated texts by three independent raters, and its results were compared with an automatic translation metric.

The final model comprises five levels: lexical-onomastic, cognitive-conceptual, pragmatic-discursive, narrative-structural and, optionally, multimodal. These levels contain eleven dimensions, each scored on a four-point scale from "lost or distorted" to "fully preserved or appropriately adapted", together with a set of error flags such as misrecognition, Westernisation, exoticisation, anachronism, calquing and value substitution. Scores are combined into a weighted Linguocultural Adequacy Index. The pilot study showed that the model can be applied reliably and that it captures cultural problems that automatic metrics miss. The article discusses the model's theoretical basis, its use by teachers, translators, publishers and AI developers, and its limitations.

Keywords: linguocultural adequacy, evaluation model, AI-generated texts, folklore, fairy tales, translation quality assessment, cultural bias, large language models, English, Uzbek.

Аннотация

Большие языковые модели (ChatGPT, Claude, Gemini и др.) сегодня за секунды пересказывают, переводят и сочиняют народные сказки, и их тексты уже используются на уроках, в детских книгах и в интернете. Однако общепринятого способа оценить, насколько такой текст культурно адекватен, не существует. Стандартные метрики качества машинного перевода, такие как BLEU, или общие оценки беглости показывают, грамматичен ли текст и близок ли он к эталону, но не показывают, остался ли Дев Девом, осталось ли *сабр* Зумрад добродетелью и приветствует ли герой чудовище словом *бобо*. Исследования фольклорных текстов, созданных ИИ, показали, что культурные потери происходят на многих уровнях — от имён и идиом до

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

метафор, ценностей, эмоций, речевого этикета, нарративной структуры и визуальных образов, — однако эти результаты до сих пор не были сведены в практический инструмент оценки.

В статье предлагается и валидируется **Модель оценки лингвокультурной адекватности (LCAAM)** фольклорных текстов, созданных ИИ. Модель разрабатывалась в три этапа. Сначала параметры и индикаторы были выведены из лингвокультурологической, когнитивной, прагматической и нарратологической теории, а также из серии эмпирических исследований английских и узбекских сказок, пересказанных ИИ. Затем модель была уточнена и валидирована экспертной группой в ходе двухраундовой процедуры Дельфи с использованием индексов содержательной валидности. Наконец, модель была апробирована на выборке текстов ИИ тремя независимыми оценщиками, а её результаты сопоставлены с автоматической метрикой перевода.

Итоговая модель включает пять уровней: лексико-ономастический, когнитивно-концептуальный, прагматико-дискурсивный, нарративно-структурный и, факультативно, мультимодальный. Эти уровни содержат одиннадцать параметров, каждый из которых оценивается по четырёхбалльной шкале — от «утрачено или искажено» до «полностью сохранено или уместно адаптировано», — а также набор маркеров ошибок: неверное распознавание, вестернизация, экзотизация, анахронизм, калькирование и подмена ценностей. Баллы сводятся во взвешенный Индекс лингвокультурной адекватности. Апробация показала, что модель применяется надёжно и выявляет культурные проблемы, которые автоматические метрики не улавливают. Обсуждаются теоретические основания модели, её использование учителями, переводчиками, издателями и разработчиками ИИ, а также её ограничения.

Ключевые слова: лингвокультурная адекватность, модель оценки, тексты, созданные ИИ, фольклор, сказки, оценка качества перевода, культурная предвзятость, большие языковые модели, английский язык, узбекский язык.

Annotatsiya

ChatGPT, Claude, Gemini kabi katta til modellari bugun xalq ertaklarini bir necha soniyada qayta hikoya qilmoqda, tarjima qilmoqda va to’qimoqda, ularning matnlari esa allaqachon darslarda, bolalar kitoblarida va internetda qo’llanilmoqda. Biroq bunday matnning madaniy jihatdan adekvat yoki adekvat emasligini baholashning umume’tirof etilgan usuli mavjud emas. BLEU kabi mashina tarjimasi sifatini o’lchovchi standart ko’rsatkichlar yoki umumiy ravonlik baholari matn grammatik jihatdan to’g’ri va namunaga yaqinligini ko’rsata oladi, lekin Dev hamon Dev bo’lib qolganini, Zumradning *sabri* hamon fazilat ekanini yoki qahramon devga hamon «bobo» deb murojaat qilishini ko’rsata olmaydi. Sun’iy intellekt yaratgan folklor matnlari bo’yicha tadqiqotlar madaniy yo’qotishlar ko’plab darajalarda — ismlar va iboralardan tortib metaforalar, qadriyatlar, his-tuyg’ular, nutq etiketi, hikoyaviy tuzilish va vizual obrazlargacha — yuz berishini ko’rsatdi, ammo bu natijalar hali amaliy baholash vositasiga birlashtirilmagan.

Maqolada sun’iy intellekt yaratgan folklor matnlari uchun **Lingvokulturologik adekvatlikni baholash modeli (LCAAM)** taklif etiladi va asoslanadi. Model uch bosqichda ishlab chiqildi. Birinchidan, o’lchovlar va ko’rsatkichlar lingvokulturologik, kognitiv, pragmatik va narratologik nazariyalardan hamda sun’iy intellekt qayta hikoya qilgan ingliz va o’zbek ertaklariga oid bir qator empirik tadqiqotlardan kelib chiqib belgilandi. Ikkinchidan, model ekspertlar guruhi tomonidan ikki bosqichli Delfi usuli va mazmuniy validlik indekslari yordamida takomillashtirildi va tasdiqlandi. Uchinchidan, model sun’iy intellekt matnlari namunasida uchta mustaqil baholovchi tomonidan sinovdan o’tkazildi va uning natijalari avtomatik tarjima ko’rsatkichi bilan solishtirildi.

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

Yakuniy model besh darajadan iborat: leksik-onomastik, kognitiv-konseptual, pragmatik-diskursiv, hikoyaviy-tuzilmaviy va ixtiyoriy ravishda multimodal. Bu darajalar o'n bitta o'lchovni o'z ichiga oladi. Ularning har biri «yo'qolgan yoki buzilgan»dan «to'liq saqlangan yoki o'rinli moslashtirilgan»gacha bo'lgan to'rt ballik shkala bo'yicha baholanadi. Bundan tashqari, noto'g'ri tanib olish, g'arblashtirish, ekzotiklashtirish, anaxronizm, kalka va qadriyat almashinuvi kabi xato belgilari qayd etiladi. Ballar vaznli Lingvokulturologik adekvatlik indeksiga birlashtiriladi. Sinov tadqiqoti model ishonchli qo'llanilishini va avtomatik ko'rsatkichlar payqamaydigan madaniy muammolarni aniqlashini ko'rsatdi. Maqolada modelning nazariy asoslari, undan o'qituvchilar, tarjimonlar, nashriyotlar va sun'iy intellekt ishlab chiquvchilari tomonidan foydalanish imkoniyatlari hamda uning cheklovlari muhokama qilinadi.

Kalit so'zlar: lingvokulturologik adekvatlik, baholash modeli, sun'iy intellekt yaratgan matnlar, folklor, ertaklar, tarjima sifatini baholash, madaniy moyillik, katta til modellari, ingliz tili, o'zbek tili.

Introduction

Imagine a teacher in a pedagogical institute in Uzbekistan who wants to use an English version of *Zumrad va Qimmat* in her class. She asks an AI model to translate the tale, and within seconds she receives a fluent, grammatically perfect text. But is it a good text? The English is correct, yet Zumrad's name no longer means "emerald", her *sabr* has become ordinary *patience*, the old woman in the forest has turned into a kind counsellor who tells her that "her feelings are valid", the closing formula of the forty-day feast has disappeared, and the tale now ends with an explicit lesson about believing in oneself. None of these changes would be noticed by a spell-checker, a grammar tool or a standard machine translation metric. All of them matter for the cultural meaning of the tale.

This scenario illustrates a growing problem. Large language models are rapidly becoming storytellers, translators and adapters of folklore. They are used to create reading materials for language learners, to produce children's books and to make traditional tales accessible to new audiences. Research has shown, however, that LLMs have a strong cultural bias towards Western, and especially Anglo-American, content and values (Naous et al., 2024; Tao et al., 2024; AlKhamissi et al., 2024), and that they perform less well in low-resource languages such as Uzbek (Robinson et al., 2023). In the case of folklore, this bias affects every level of the text.

A series of studies within the present research project has examined these effects in detail, comparing canonical English and Uzbek fairy tales with AI-generated retellings, translations and adaptations. These studies found that AI tends to transliterate meaningful names without explanation; to replace mythological beings such as the Dev and the Pari with Western counterparts; to preserve universal metaphors but flatten culture-specific ones; to narrow and individualise value concepts such as *sabr*, *mardlik* and *or-nomus*; to psychologise emotions; to shift speech etiquette from kinship and hierarchy towards Anglo-American politeness; to reshape the Proppian structure of the tale into a modern "hero's journey"; to replace the Uzbek wolf with the Big Bad Wolf; and, in image generation, to dissolve Uzbek heroes into a generic "Oriental" style. Taken together, these findings show that the cultural adequacy of an AI-generated folklore text is a multidimensional phenomenon.

Yet there is still no practical tool for assessing it. Existing approaches to translation quality assessment focus on accuracy, fluency and terminology (House, 2015; Lommel et al., 2014), while automatic metrics such as BLEU measure surface similarity to a reference translation (Papineni et al., 2002). Research on the cultural evaluation of LLMs has mainly used survey-style questions about values and knowledge rather than the analysis of culturally rich texts (Adilazuarda et al., 2024). Teachers, translators, publishers and developers therefore lack a systematic way to answer a simple question: *Is this AI-generated tale culturally adequate?*

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

The aim of this article is to propose and validate a model for assessing the linguocultural adequacy of AI-generated folklore texts. The study addresses three research questions:

1. Which dimensions and indicators should a model of linguocultural adequacy for AI-generated folklore texts include?
2. Do experts consider these dimensions relevant and clear, and how should they be weighted?
3. Can the model be applied reliably by independent raters, and does it capture aspects of quality that automatic metrics miss?

The article contributes to linguoculturology, translation studies and the evaluation of AI by integrating findings from several levels of linguistic and cultural analysis into a single, validated assessment tool. Its practical value lies in providing teachers, translators, publishers and developers with a usable instrument and checklist.

Literature Review

Adequacy in translation theory

The notion of **adequacy** has a long history in translation studies. Toury (1995) contrasted *adequacy*, meaning faithfulness to the norms of the source text and culture, with *acceptability*, meaning conformity to the norms of the target culture, and argued that every translation negotiates between the two. Reiss and Vermeer (1984/2013), in their *skopos* theory, defined adequacy in relation to the purpose of the translation: a translation is adequate if it fulfils its intended function for its intended audience. Nida's (1964) distinction between formal and dynamic equivalence and Venuti's (1995) distinction between domestication and foreignisation describe related tensions. For folklore texts, adequacy implies not only accurate transfer of meaning but also preservation, or appropriate adaptation, of the cultural world of the text.

Translation quality assessment

Systematic models of translation quality assessment have been developed by House (2015), whose functional-pragmatic model compares the source and target texts in terms of register and genre, and by the Multidimensional Quality Metrics (MQM) framework, which classifies translation errors into categories such as accuracy, fluency, terminology and locale convention (Lommel et al., 2014). Freitag et al. (2021) showed that expert evaluation based on MQM is more reliable than crowd-sourced ratings and that automatic metrics often disagree with experts. MQM includes a category for "locale convention", which covers dates, currencies and similar items, but it does not address deeper cultural meanings such as metaphors, values or narrative structure.

Automatic evaluation and its limits

Automatic metrics such as BLEU (Papineni et al., 2002) measure the overlap of words and phrases between a machine output and a reference translation. They are fast and widely used, but they are blind to cultural meaning: a translation that renders *Dev bobo* as "Mr Giant" may score well if the reference uses similar words, while a culturally careful translation with an explanatory gloss may score poorly. Moreover, for retellings and adaptations, which are common uses of LLMs, there is often no single reference text at all.

Cultural evaluation of LLMs

Research on the cultural dimension of LLMs has grown rapidly. Hershcovich et al. (2022) outlined the challenges of cross-cultural natural language processing, distinguishing linguistic form, common ground, aboutness and values. Adilazuarda et al. (2024) surveyed studies that measure "culture" in LLMs and found that most use proxies such as survey questions or factual knowledge, with little attention to culturally rich discourse. Studies of cultural alignment (Cao et al., 2023; AlKhamissi et al., 2024; Tao et

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

al., 2024) and cultural bias (Naous et al., 2024) have demonstrated a Western bias but have not proposed tools for evaluating texts such as folk tales.

Linguoculturological foundations

Linguoculturology offers a rich theoretical basis for analysing cultural meaning in texts. Key notions include the **linguocultureme** (Vorob'ev, 1997), **cultural codes** (Krasnykh, 2002), the **linguocultural concept** with its notional, figurative and value layers (Karasik, 2002), **cultural key words** (Wierzbicka, 1997) and **cultural conceptualisations** (Sharifian, 2017). Cognitive linguistics adds the analysis of conceptual metaphors and cultural models (Lakoff & Johnson, 1980; Kövecses, 2005), pragmatics adds politeness and address (Brown & Levinson, 1987), and folkloristics adds the morphology and style of the folk tale (Propp, 1968; Lüthi, 1982). Each of these perspectives illuminates one level of cultural adequacy; together they provide the building blocks of a comprehensive model.

Research gap

There are models of translation quality and studies of cultural bias in AI, but no model that integrates linguocultural, cognitive, pragmatic and narrative dimensions to evaluate AI-generated folklore texts. This article aims to fill that gap.

Methods

Research design

The study follows a three-phase design for developing and validating an assessment instrument: (1) theoretical and empirical derivation of dimensions; (2) expert validation through a Delphi procedure; and (3) a pilot application to test reliability and usefulness.

Phase 1: Derivation of dimensions

Dimensions and indicators were derived from two sources. First, the theoretical literature reviewed above provided the conceptual foundation for each level of analysis. Second, the findings of a series of empirical studies within this research project, covering national-cultural codes, conceptual metaphors, gender representation, mythological characters, speech etiquette, value concepts, visual representation, zoomorphic characters, names, narrative structure and emotions in AI-generated English and Uzbek fairy tales, were systematically reviewed. For each study, the main types of cultural loss or distortion were listed and grouped. Similar categories were merged, and each resulting dimension was given a definition, indicators and scoring descriptors. This produced a preliminary model with twelve dimensions.

Phase 2: Expert validation

A panel of experts was formed, including linguists specialising in linguoculturology and cognitive linguistics, folklorists, professional translators between Uzbek and English, and teachers of English at pedagogical institutions. The experts were selected for their expertise in at least one relevant area and their knowledge of both Uzbek and English. The validation followed a two-round **Delphi procedure** (Hsu & Sandford, 2007).

In the first round, the experts rated each dimension for **relevance** on a four-point scale (1 = not relevant, 4 = highly relevant) and commented on its definition, indicators and scoring descriptors. The item-level content validity index (I-CVI) was calculated as the proportion of experts rating a dimension 3 or 4, and the scale-level index (S-CVI/Ave) as the average of the I-CVIs (Lynn, 1986; Polit & Beck, 2006). Dimensions with an I-CVI below 0.78 were revised or removed. In the second round, the experts reviewed the revised model, re-rated relevance and clarity, and assigned **weights** to the five levels by distributing 100 points among them. The final weights were calculated as the mean of the expert allocations.

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

Phase 3: Pilot application

The final model was applied to a pilot sample of 30 AI-generated texts, drawn from the corpora of the earlier studies, which included retellings, translations and modernisations of English and Uzbek tales produced by three LLMs. Three trained raters independently scored each text on all dimensions. Inter-rater reliability was assessed with Krippendorff's alpha for ordinal data (Krippendorff, 2018). To test whether the model captures aspects of quality not reflected in automatic evaluation, the Linguocultural Adequacy Index of each translated text was compared with its BLEU score against a human reference translation (Papineni et al., 2002). The raters also gave qualitative feedback on the usability of the model.

Results.

1. The final model: structure

After the Delphi procedure, the preliminary twelve dimensions were reduced to **eleven**, grouped into **five levels**. Two closely related dimensions concerning character images were merged, and the definitions of several dimensions were clarified in response to expert comments.

Level	Dimension	What is assessed	Theoretical basis
I. Lexical-onomastic	1. Culture-specific lexis and realia	Food, clothing, places, objects, kinship terms	Vorob'ev (1997); Krasnykh (2002)
	2. Proper names	Meaning, connotation and function of character names	Literary onomastics; Hermans (1988)
II. Cognitive-conceptual	3. Conceptual metaphors and imagery	Culture-specific metaphors and idioms	Lakoff & Johnson (1980); Kövecses (2005)
	4. Value concepts	Notional, figurative and value layers of key values	Karasik (2002); Wierzbicka (1997)
	5. Emotion concepts	Mode, intensity and cultural form of emotional expression	Wierzbicka (1999); Kövecses (2000)
III. Pragmatic-discursive	6. Speech etiquette	Address forms, honorifics, politeness strategies	Brown & Levinson (1987); Ide (1989)
	7. Formulas and paremias	Opening and closing formulas, proverbs, blessings	Folkloristics; paremiology
IV. Narrative-structural	8. Narrative structure	Proppian functions, triplication, chronotopes, endings	Propp (1968); Olrik (1909/1965)
	9. Character images	Heroes, mythological beings, animals, gender roles	Folkloristics; cultural codes
	10. Style and genre	Folk-tale style versus modern psychological narrative	Lüthi (1982)
V. Multimodal (optional)	11. Visual cultural codes	Clothing, setting, artefacts, ornament, physiognomy in images	Kress & van Leeuwen (2006)

2. Scoring scale

Each dimension is scored on a four-point scale. The general descriptors are as follows:

Score	Label	General descriptor
3	Fully preserved or	The cultural element is preserved, or adapted in a way that keeps

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

Score	Label	General descriptor
	appropriately adapted	its meaning and function (e.g., kept with a gloss)
2	Partially preserved or neutralised	The core meaning is kept but cultural specificity is reduced (e.g., *otajon → father*)
1	Substituted	The element is replaced by a counterpart from another culture (e.g., *Pari → fairy*; *sabr → self-assertion*)
0	Lost or distorted	The element is omitted, misrecognised or its meaning is reversed (e.g., *bo'riday* translated with predatory connotations)
N/A	Not applicable	The element does not occur in the source text

Each dimension also has specific descriptors. For example, for **dimension 2 (proper names)**, a score of 3 means that a speaking name is kept and its meaning made clear (*Zumrad, "Emerald"*), 2 means that the name is kept without explanation, 1 means that it is replaced by a target-culture name (*Zumrad → Emily*), and 0 means that it is misread or its function is lost (*Kal* read as *Cal*). For **dimension 8 (narrative structure)**, 3 means that the key Proppian functions, triplication and traditional formulas are preserved, 2 means that the core sequence is kept but some functions or repetitions are lost, 1 means that a modern template such as an inner arc, a forgiven villain or an explicit moral replaces the folk structure, and 0 means that the structure is fundamentally altered.

3. Error flags

In addition to the scores, raters record the presence of specific **error flags**, which identify the type of cultural problem and help users understand and correct it:

Error flag	Definition	Example
Misrecognition	A culture-specific item is interpreted through an unrelated meaning	*Dev* rendered as a Hindu deity; *Kal* as the name *Cal*
Westernisation	A local element is replaced by a Western counterpart	*Pari* → winged *fairy*; *palov* → *pie*
Exoticisation	Culture is represented through exaggerated "Oriental" stereotypes	Generic domes, deserts, veils
Anachronism	An element from the wrong period or setting is introduced	Registan madrasas in a nomadic epic
Calque	A target-language structure is unnaturally copied into the source language	Overuse of *his qildi* in Uzbek
Value substitution	A cultural value is replaced by another or its evaluation is reversed	*Sabr* presented as weakness
Psychologisation	External, embodied or action-based expression is replaced by introspection	*Yuragi orqasiga tortib ketdi → she felt anxious*
Over-moralisation	An explicit moral not present in the source is added	"And so she learned that…"

The pilot application showed that the model could be applied by trained raters with acceptable reliability. Agreement was highest for the more concrete dimensions, such as names, realia and formulas, and somewhat lower for the more interpretive dimensions, such as value concepts and style, which indicates where rater training is most needed. AI texts based on Uzbek sources received lower LCAI scores than texts based on English sources, which is consistent with the findings of the earlier studies. The correlation between LCAI and BLEU was weak, confirming that automatic metrics do not capture linguocultural adequacy: several fluent translations with relatively high BLEU scores received low LCAI

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

scores because of value substitution, the loss of names and formulas, or the Westernisation of mythological beings.

7. An illustrative application

To show how the model works in practice, the following table presents an **illustrative** scoring of a typical AI-generated English translation of *Zumrad va Qimmat*, based on the types of changes documented in the earlier studies:

Dimension	Score	Comment	Error flags
1. Culture-specific lexis and realia	2	*Sandiq* rendered as *chest*; food and clothing generalised	—
2. Proper names	2	*Zumrad* and *Qimmat* kept without explanation; contrast in meaning lost	—
3. Conceptual metaphors and imagery	1	Somatic idioms replaced with abstract terms	Psychologisation
4. Value concepts	1	*Sabr* reduced to *patience*; religious layer lost	Value substitution
5. Emotion concepts	1	Emotions named and analysed; hyperbole moderated	Psychologisation
6. Speech etiquette	2	Respectful address to the old woman partly kept; blessings omitted	—
7. Formulas and paremias	1	Opening replaced with *"Once upon a time"*; closing formula lost	Westernisation
8. Narrative structure	2	Core sequence kept; donor test shortened	—
9. Character images	2	Old woman kept but becomes a "counsellor"	—
10. Style and genre	1	Explicit moral added at the end	Over-moralisation

In this example, the text would receive an LCAI in the band "substantial revision required". Its translation would nevertheless be grammatically correct and fluent. The error flags point directly to the revisions needed: restoring the formulas, glossing the names, re-establishing the value of *sabr* and showing emotions through action.

Discussion

From diagnosis to assessment

The earlier studies in this project diagnosed specific cultural losses in AI-generated folklore. The model proposed here turns these diagnoses into an **assessment tool**. Its main contribution is integrative: it brings together the lexical, cognitive, pragmatic, narrative and visual levels of cultural meaning, each grounded in an established theoretical tradition, within a single framework with clear descriptors and a composite index. In doing so, it operationalises the notion of linguocultural adequacy, which has been widely used in theory but rarely measured in practice.

Adequacy, not literalism

It is important to emphasise that the model does not demand literal preservation of every cultural element. In line with skopos theory (Reiss & Vermeer, 1984/2013) and Toury's (1995) notion of adequacy, the highest score is given both to preservation and to **appropriate adaptation**: an adaptation that keeps the meaning and function of the cultural element, such as a name with a gloss or an idiom rendered with a comparable image. What the model penalises is not change as such, but change that

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

erases, substitutes or distorts the source culture without the user's awareness. This makes the model compatible with different purposes, from scholarly translation to adaptation for young readers.

Why automatic metrics are not enough

The weak correlation between LCAI and BLEU in the pilot study confirms what translation scholars have long argued: fluency and surface accuracy are not the same as quality (House, 2015; Freitag et al., 2021). For culturally rich texts such as folklore, the most serious problems are often invisible at the surface: a well-formed sentence may contain a misrecognised mythological being, a reversed value or a lost blessing. The model therefore complements, rather than replaces, existing metrics, and it fills a gap in current approaches to the cultural evaluation of LLMs, which rarely examine culturally rich discourse (Adilazuarda et al., 2024).

Uses of the model

The model can be used in several ways. **Teachers** can use a simplified version as a checklist when selecting or adapting AI-generated texts for class, and as a tool for developing students' intercultural awareness: asking students to score an AI tale encourages them to notice cultural meanings. **Translators and editors** can use the error flags to identify and correct problems systematically. **Publishers** can adopt minimum LCAI thresholds for AI-assisted folklore publications. **AI developers** can use the model as a basis for culturally informed benchmarks and for improving models' performance on low-resource cultures. Finally, the model's dimensions can be turned into **prompt guidelines**: for example, "Keep traditional names and explain their meaning", "Preserve opening and closing formulas", "Show emotions through action" and "Do not add an explicit moral".

Transferability

Although the model was developed on English and Uzbek material, its structure is not language-specific. The five levels and eleven dimensions describe general aspects of cultural meaning in folklore, while the specific descriptors and examples can be adapted to other languages and traditions. It may be particularly useful for other Turkic and Central Asian cultures, which share many features with Uzbek folklore and face similar risks of underrepresentation in AI training data.

Limitations

The model has several limitations. First, linguocultural assessment inevitably involves expert judgement; despite clear descriptors and rater training, some dimensions, especially value concepts and style, remain partly subjective. Second, the expert panel and the pilot sample were limited in size, and the weights reflect the views of a particular group of experts. Third, the model is time-consuming to apply in full, which may limit its use in large-scale evaluation; future work could explore semi-automatic support, for example by using LLMs to pre-annotate texts for human raters. Finally, the model assesses texts, not their effects; future research should examine how texts with different LCAI scores are perceived by readers, especially children.

Conclusion

As AI becomes a storyteller, we need ways to judge not only whether its stories are fluent but also whether they are true to the cultures they come from. This article has proposed and validated a model for assessing the linguocultural adequacy of AI-generated folklore texts.

The main results can be summarised as follows:

1. The Linguocultural Adequacy Assessment Model (LCAAM) comprises five levels (lexical-onomastic, cognitive-conceptual, pragmatic-discursive, narrative-structural and multimodal) and eleven dimensions, each grounded in established linguistic, cultural and folkloristic theory.

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

2. Each dimension is scored on a four-point scale, supported by specific descriptors and eight error flags that identify typical problems, such as misrecognition, Westernisation, exoticisation, anachronism, calquing, value substitution, psychologisation and over-moralisation.
3. Scores are combined into a weighted Linguocultural Adequacy Index, with weights derived from expert judgement and interpretation bands for practical use.
4. Expert validation through a Delphi procedure confirmed the relevance and clarity of the dimensions and led to refinements of the model.
5. A pilot application showed that the model can be applied reliably and that it captures cultural problems that automatic metrics such as BLEU do not detect.

The model is offered as a practical tool for teachers, translators, publishers and AI developers, and as a contribution to the theory of linguocultural adequacy. Its underlying message is simple: a fairy tale retold by AI should be judged not only by how well it reads, but also by how well it remembers where it came from.

References

1. Adilazuarda, M. F., Mukherjee, S., Lavania, P., Singh, S., Aji, A. F., O'Neill, J., Modi, A., & Choudhury, M. (2024). Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 15763–15784).
2. AlKhamissi, B., ElNokrashy, M., AlKhamissi, M., & Diab, M. (2024). Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 12404–12422).
3. Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.
4. Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)* (pp. 53–67).
5. 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.
6. Hermans, T. (1988). On translating proper names, with reference to *De Witte* and *Max Havelaar*. In M. J. Wintle (Ed.), *Modern Dutch studies* (pp. 11–24). London: Athlone Press.
7. Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., & Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 6997–7013).
8. House, J. (2015). *Translation quality assessment: Past and present*. London: Routledge.
9. Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessment, Research, and Evaluation, 12*(10), 1–8.
10. Ide, S. (1989). Formal forms and discernment: Two neglected aspects of universals of linguistic politeness. *Multilingua, 8*(2–3), 223–248.
11. Karasik, V. I. (2002). *Yazykovoy krug: lichnost', kontsepty, diskurs* [The language circle: Personality, concepts, discourse]. Volgograd: Peremena.
12. Kövecses, Z. (2000). *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge: Cambridge University Press.

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.

13. Kövecses, Z. (2005). *Metaphor in culture: Universality and variation*. Cambridge: Cambridge University Press.
14. Krasnykh, V. V. (2002). *Etnopsikholingvistika i lingvokul'turologiya* [Ethnopsycholinguistics and linguoculturology]. Moscow: Gnozis.
15. Kress, G., & van Leeuwen, T. (2006). *Reading images: The grammar of visual design* (2nd ed.). London: Routledge.
16. Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Thousand Oaks, CA: SAGE.
17. Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.
18. Lommel, A., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica, 12*, 455–463.
19. Lüthi, M. (1982). *The European folktale: Form and nature* (J. D. Niles, Trans.). Philadelphia: Institute for the Study of Human Issues.
20. Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385.
21. Naous, T., Ryan, M. J., Ritter, A., & Xu, W. (2024). Having beer after prayer? Measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 16366–16393).
22. Nida, E. A. (1964). *Toward a science of translating*. Leiden: E. J. Brill.
23. Olrik, A. (1965). Epic laws of folk narrative. In A. Dundes (Ed.), *The study of folklore* (pp. 129–141). Englewood Cliffs, NJ: Prentice-Hall. (Original work published 1909)
24. 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).
25. Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497.
26. Propp, V. (1968). *Morphology of the folktale* (L. Scott, Trans.; 2nd ed.). Austin: University of Texas Press. (Original work published 1928)
27. Reiss, K., & Vermeer, H. J. (2013). *Towards a general theory of translational action: Skopos theory explained* (C. Nord, Trans.). London: Routledge. (Original work published 1984)
28. Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation* (pp. 392–418).
29. Sharifian, F. (2017). *Cultural linguistics: Cultural conceptualisations and language*. Amsterdam: John Benjamins.
30. Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus, 3*(9), pgae346. https://doi.org/10.1093/pnasnexus/pgae346
31. Toury, G. (1995). *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.
32. Venuti, L. (1995). *The translator's invisibility: A history of translation*. London: Routledge.
33. Vorob'ev, V. V. (1997). *Lingvokul'turologiya: teoriya i metody* [Linguoculturology: Theory and methods]. Moscow: RUDN Publishing.
34. Wierzbicka, A. (1997). *Understanding cultures through their key words: English, Russian, Polish, German, and Japanese*. New York: Oxford University Press.

© 2026 London International Monthly Conference on Multidisciplinary Research and Innovation. All Rights Reserved.