International Journal of Language and Literary Studies
Volume 8, Issue 6, 2026
Homepage : http://ijlls.org/index.php/ijlls

The Effect of Prompts on LLMs Performance in Translating Maribi Proverbs from Arabic into English: An Analysis of Strategies and Accuracy

Othman Saleh Mahdy Mohammed
Assistant Professor, Department of English, Faculty of Arts and Humanities, University of Saba Region, Yemen
othmanwosabi924@gmail.com

DOI: http://doi.org/ 10.36892/ijlls.v8i6.2863
APA Citation: Mohammed, O. S. M. (2026). The Effect of Prompts on LLMs Performance in Translating Maribi Proverbs from Arabic into English: An Analysis of Strategies and Accuracy. *International Journal of Language and Literary Studies. 8*(6).318-334. http://doi.org/10.36892/ijlls.v8i6.2863

Received: *18/08/2026*

Accepted: *16/9/2026*

Keywords: *LLMs, Prompt, Maribi, Proverbs, Strategies, Accuracy.*

Abstract

This study aims to examine the effect of employing prompts on the accuracy and strategies used by LLMs in translating Maribi proverbs. To achieve the objectives of this study, 74 Maribi proverbs were processed into Chat GPT, Gemini and DeepSeek. A quantitative approach was utilized and the frequencies, percentages, and means were run via SPSS to yield the results. Based on the outcomes of these Models, a thorough analysis of the accuracy of translation and the strategies used was made. The findings of the study revealed that the prompts had a great effect on all LLMs-particularly Gemini. Also, the strategies were altered from literal to paraphrase and the use of proverbs of similar meanings and form. The study concluded with implications for teaching translation courses, stressing on teaching students how to use and design prompts.

1. INTRODUCTION

The rapid incorporation of Large Language Models (henceforth LLMs) into the translation realm has fundamentally changed the tasks and the ways of translation, as LLMs offer rapid and accurate (to some extent) translations. However, LLMs are under scrutiny for handling culture-bound idiomatic language. Proverbs, which rely heavily on implicit cultural knowledge and dialectal hints, require moving beyond literal linguistic equivalence to achieve authentic pragmatic tone. One suffice definition to be mentioned here in this study by Newmark (1988) stating that "translation is rendering the meaning of a text into another language in the way that the author intended the text" (p. 5). Translation is a challenging craft that involves transferring meaning and providing a natural equivalent for cultural and lexical differences between the source language (SL) and target language(TL). Culture and translation are inseparable (Lau et al., 2022; Jibreel, 2023). While translating, it is very important to understand the two languages scrupulously; translating the cultural differences between them further complicates the task of translation. LLMs application are on the focus to determine whether such application can compromise between the two inseparable elements.

Marib Governorate, located in Yemen, possesses a rich oral tradition where such proverbs serve as vibrant containers of historical, social, and cultural beliefs. Translating these culturally specific items (CSIs) from the local Yemeni jargon into English is a daunting task that provokes stressing demands of evaluating the pragmatic competence of LLMs. Although artificial

intelligence (AI) has revolutionised automated translation, its proficiency in decoding and transferring highly localised, dialect-bound items remains largely unexplored. To address this gap, several studies have examined the direct use of LLMs in translating CSIs and to check their abilities in accurate rendition; this study takes a different dimension as it assesses the effect of prompts on LLMs in translating Maribi proverbs from Arabic into English by systematically analysing the translation strategies employed and evaluating the semantic and cultural accuracy of the generated translations before and after using prompts for three prominent LLMs.

1.1.Statement of the problem

Translating proverbs from Arabic into English poses serious problems for translators (let them be humans or machines) due to the linguistic, stylistic, and cultural gaps between English and Arabic. The situation is getting more complicated when it comes to the task of rendering CSIs, which are furthermore rooted in the dialectical items- proverbs of the Marib region, where there is limited interaction with many other dialects. The translators- let them be called here AI users- translate these CSIs immediately using such LLMs without designing prompts. The study is limited to investigating the LLMs Baker's (1992) model of translating proverbs and idioms has been adopted for this study since it provides an appropriate framework for the translation of proverbs.

1.2.Research Objectives

This research aims to:

1- Figure out the strategies used by LLMs in translating Maribi proverbs into the English language before and after using prompts.

2- Scrutinise the extent of the effect of using prompts on the accuracy of LLMs in translating Maribi proverbs into the English language.

3- Identify the most accurate translation among the three LLMs and which one benefited more from the prompt.

2. LITERATURE REVIEW

The appearance of LLMs has made a great transformation in the field of translation, as LLMs have reshaped both theoretical views and professional practices (Ataman et al., 2025; Wang & Wu, 2026). LLMs, developed by OpenAI, launched in 2022 with key distinctions (Kukreja et al., 2024) which evidenced the true beginning of LLMs (Ray, 2023). Unlike earlier neural systems that struggled with rigid phrase-based translation and were trained specifically for translation, LLMs offer greater flexibility in style, register, and adaptability to different domains, are trained on massive multilingual and multi-domain data and can perform translation as part of their broader text generation abilities (Ataman et al., 2025; Gain et al., 2026). These models show strengths in producing fluent and context-aware translations, sometimes handling discourse and style better than traditional neural machine translation. Because of their flexibility and integration into many applications, LLMs now represent the latest stage in MT evolution.

2.1. Large Language Models

AI really came into its own in the field of translation in 2017 with DeepL. This neural translation software is based on data from the translation website Linguee, a multilingual dictionary that compares translations in nearly 20 languages. This marked the beginning of neural machine translation. This intelligent technology, which is based on artificial neurons, considers the entire text and its context. Furthermore, neural machine translation is able to continuously improve and refine itself, thanks to the flow of data it receives. Neural machine translation is a major technological advance that allows us to obtain a text that is much more qualitative and fluid than with machine translation. LLMs represent a definitive departure from traditional, rule-based machine translation paradigms (Ahmed, 2025). LLMs have verified remarkable competencies in producing human-like translations (Mahdy et al., 2020; Al-Wasy & Mohammed, 2024; Jiang et al., 2024 Qian & Kong, 2024; Noman et al, 2025). Instead of relying on static bilingual dictionaries, these LLMs utilise transformer-based neural networks trained on massive input sequences, with language models using between 500 and 2000 tokens in length before pre-training (Kumar, 2024). Despite their many advantages, LLMs are far from reliable and still have many limitations in the field of translation (Wang et al., 2024). Indeed, this technology is not able to adapt the translation to the target readership. Nor can it consider local cultural norms (Budimir, 2025) and customs, the clients' expectations, the style, and the translation's intention. These are essential elements in translation to obtain texts that are respectful of the local culture, adapted to the target audience and faithful to the source text. Furthermore, while AI translation can be effective for the cognate languages (English, French, Spanish, German, Dutch, Italian, etc.), it is much less effective for non-cognate languages or dialects for which little data exists.

2.2.Proverbs Translation and Culture

According to Karasüleymanoğlu (2002), culture is defined as "all the material and spiritual values that emerge in the process of socialisation and the impact that humans create on the environment while applying them, as well as all the tools that show the extent of their sovereignty" (Karasüleymanoğlu, 2002, as cited in Ertugruloglu et al., 2024, p. 132). Cultural universals are found in all human societies; these include expressive forms like art, music, dance, ritual, religion, and technologies like tool usage, cooking, shelter, and clothing. According to Villers (2022), a proverb is "a short, generally known sentence of the folk which contains wisdom, truth, morals and traditional views in a metaphorical, fixed and memorizable form and which is handed down from generation to generation." (p.3).

Translation of proverbs is not an easy task since it requires enough linguistic and cultural knowledge in both languages. With relation to cultural aspect and according to House (2009) as cited in Al Rousan and Shatnawi (2023) reported that "translating is not only a linguistic act, it is also a cultural one". (p. 448). In this framework, House stresses the importance of paying enough attention to cultural aspects in translation since proverbs contain cultural and linguistic features.

As for linguistic aspects, it's clear that translating proverbs requires linguistic knowledge in relation to finding the correct translation equivalence. In this framework, Nida (1964) proposes

two translation methods, namely: formal equivalence and dynamic equivalence. Formal equivalence in which the translator focuses attention on the form and content of the SL message, and the dynamic equivalence in which the translator focuses primarily on providing an equivalent effect for the SL message.

Maribi proverbs are not mere phrases but cultural palimpsests. They encode layers of meaning tied to Yemen's pastoral heritage, where cattle symbolise both livelihood and social standing. A literal rendering risks reducing such proverbs to absurdity, stripping them of their metaphorical weight and historical resonance. The challenge intensifies when these sayings intersect with themes of resilience, social critique, or communal ethics—concepts deeply rooted in Maribi tribal code but often alien to English-speaking audiences.

2.3.Yemeni Proverbs

Yemeni proverbs function as important cultural artefacts to demonstrate the structural elements of social traditions and historical events which define Yemeni gender relations. These proverbs contain wisdom collected from communities through which societal norms emerge, as they show gender power dynamics, especially regarding how women face stereotypes and discrimination (al-Hussami & al-Hiba, 2023). History and local wisdom can be better understood through proverbs, according to scholar 'Abd Allāh Al-Baraddūnī, who explained their value in providing cultural interpretations of historical figures and events (Kamal, 2022).

Among Yemeni proverbs that base their themes on fauna are the agricultural lifestyle type alongside the nomadic lifestyle type, along with different thematic expressions throughout regional Yemen (Mohammed & Ho-Abdullah, 2021; Farghal & Saeed, 2022; al-Saqqaf, 2023). In the realm of intercultural communication, proverbs are seen as carriers of human values, often reflecting global discourses while sidelining local narratives, thus highlighting the tension between local and global identities in the discourse of proverbs (Hazaea, 2017). Collectively, these studies underscore the multifaceted role of Yemeni proverbs as cultural artefacts that preserve and communicate the complex tapestry of Yemeni life and values.

2.4.Challenges in Translating Maribi Proverbs

Maribi proverbs are rooted in the local dialect of Marib; when translating them into English, it becomes a very complicated task comprising linguistic and cultural issues. Maribi Arabic is a type of Yemeni dialect which is characterised by its own phonological, morphological, and syntactic features that are not similar to any standard form of English. Their meaning is not the only issue in the context of lexicography – these proverbs are also the carriers of the cultural wisdom of thousands of traditions, social norms, and personal memories of the people in central Yemen who belong to the tribal community.

The cultural dimension is probably the greatest impediment in this case. These proverbs are not just a means of conveying information, but they also perform a lot of social functions which foster communal values, or even comment on difficult topics in a very appeasing way. When a Maribi proverb talks of agricultural practices or tribal customs, it presupposes that English speakers have shared knowledge that, in most cases, English speakers lack. This is why even a linguistically correct translation still sometimes remains far off the mark (Jabak, 2018). Following the same logic, Al-Saidi (2021) points out that successful translations necessitate bicultural fluency: the grasp of both linguistic systems and their sub-cultural backgrounds.

With the emergence of AI technologies, the translation process has witnessed a radical change, so the translators- let them be professional, practicing or even users, resort to LLMs to translate the texts. So it is vitally important to check the accuracy and the strategies used by these LLMs.

2.5. Strategies for Translating Proverbs

Navigating the cultural and linguistic disproportionateness intrinsic in proverbs requires the careful application of specialised translation strategies to balance semantic fidelity with pragmatic equivalence (Shehab, 2026). To achieve this, translators typically navigate a range of strategies dictated by the SL metaphorical opaqueness and the target culture's communicative norms. At one end of this spectrum is literal translation, which preserves the exact lexical items and syntactic structures of the original; while effective for universal sayings, it often risks rendering culturally bound items-proverbs in our context- incomprehensible to the target audience. To mitigate such communicative breakdowns, the translators who deal with these proverbs are to propose a few techniques that could be successful. Among them are those such as literal translation, paraphrasing, cultural substitution, and dynamic equivalence (Bajubair &Abdulla, 2022). The translators who do this job need to be very competent because each one of these approaches has its own strengths and weaknesses; the translators may have to change their text's specifics, style, or expression.

When translators opt for the literal translation method, they want to retain the form of the source text without changing the format or style and undertaking as little structural alteration as possible (Almjlad, 2023). This approach is one which seems to be very simple but, in fact, is often far from sufficient because such expressions typically convey metaphorical and idiomatic meanings on a word-by-word basis. Basically, proverbs are the fine arts of life that satisfy the criterion of being a token of our cultural identity and contain the attributes of unique experiences, which a translation from the perspective of another language (Alfaleh, 2020).

Sometimes the result is a translation that is not only inaccurate but also fails to make any sense, most notably when cultural metaphors are dealt with, and the situations are not understood by the audience of the translated text (Al-Saidi, 2021). However, in some cases, literal translation can be more effective than other translation methods since it can preserve the universal imagery of the proverbs and, in this way, intensify the poetical and rhetorical effect of the proverb (Alharbi, 2023). Even though this method has its unique advantage, it often fails to convey the layers of feelings, underscoring the urgency of strategies like dynamic equivalence or cultural substitution to become effective cross-cultural and cross-linguistic bridges.

By rewording a phrase, the meaning in a variety of ways can be manipulated , thus, can help translating a proverb that is very difficult to be translated because it contains certain cultural elements, metaphors, or idioms that don’t have equivalents. The method’s technique is designed to translate proverbs into the local idiom and cultural norms while keeping the message constant. (Bajubair & Abdulla, 2022; Shhaiber & Haider, 2023). Paraphrasing is particularly good at expressing what someone is saying in their own words using idiomatic expressions. However, the brevity, rhythm, or poetic qualities might be lost (Alfaleh, 2020; Alharbi, 2023). In order to effectively translate, one should strike the right balance between

being clear in conveying the original text's stylistic and cultural attributes, and being faithful to the source (Saideen et al., 2024).

Dynamic equivalence aims to keep the intended significance and emotional effectiveness of the text so that the target audience can sense it just as the original one (Nida, 1964). Such a method is really timely when translating culturally-based idiomatic phrases, as it gives translators a handle to deal with metaphorical language and references that change concerning the cultural situations and practices of the target audience (Almjlad, 2023). By putting the emphasis on communicative effect, the dynamic equivalence processes convey the proverb's both emotional and pragmatic impacts, making the translation be received well by the audience (Dickins et al., 2016).

However, a strategy like this comes with a need for delicate balancing between cultural authenticity and the ease of getting the message (Saideen et al., 2024). Cheerfully, not only to linguistically but also to culturally distant or different people, but still, many people may think that dynamic equivalence is the best choice for interpretation, since it allows translators to have originality, combining the use of creativity with knowledge about the culture to use idiomatic phrases.

Baker (2018, pp. 77-87) suggested strategies that help translators when translating proverbs. These strategies are:

1- Using a proverb of similar meaning and form. In this strategy, the translator uses a target language proverb that has a meaning and lexical items similar to the source language.

2- Using a proverb of similar meaning but dissimilar form. In this strategy, the translator uses a TL proverb that has a similar meaning but different lexical form.

3- Translation by paraphrase. This strategy is used when there is no target language equivalent matching the source text proverb or the style of the target text differs from the source text.

4- Translation by omission. In this strategy, the translator deletes the whole proverb because there is no target language proverb matching the source language proverb closely. So the translator resorts to a near equivalent or even a literal translation, which is the weak version (Qasim, 2026).

3. METHODOLOGY

This study employed a quantitative research method. It focused on the frequencies and percentages of translation errors of LLMs and the most commonly used strategies by each one. The accuracies of LLMs were measured. The research explored the effectiveness of LLMs in translating Maribi proverbs and compared their performance.

3.1. Data Collection

This study collected the data from different sources- magazines, websites, and Facebook pages of Maribi people. Then 112 Proverbs were grouped. Then the researcher presented these proverbs to four native Maribi dialect speakers, aged more than 50 years, to ascertain that these proverbs related to the Maribi dialect and region, and those proverbs which were deemed to be

general were excluded. The final number of proverbs for the study was 74; they were selected according to their utilisation among Maribi people, their lively usage and their meaning being understandable by the informants, as those Maribi proverbs that were not understandable were excluded. After that, the final filtered draft of the proverbs was handed to two informants who have a bachelor's degree to check the correct written form of the proverbs. To validate the collected translation data, the annotators-two university professors evaluated the outcomes of the LLMs in both phases and checked the scaling and the strategies allotted for each proverb.

3.2.Prompts

The research design a prompt template to be processed via the selected LLMs. as the proverbs were associated with their Arabic meanings. It also required the LLMs to choose the most appropriate English equivalent. Furthermore, all LLMs were given the same prompt; the researcher did that to avoid outcome biases.

3.3. Data Analysis

The data were collected from a diverse set of Maribi proverbs from different sources. All proverbs were taken from daily speech and commonly repeated proverbs in Marib Governorate. In addition, this research used many procedures and materials for collecting the data. This study documented and analysed a total of 74 Maribi proverbs through a rigorous data collection and translation process. Three LLMs were selected: ChatGPT, Gemini and DeepSeek. After getting the results from human evaluator, the accuracy and the strategies were coded and processed via SPSS to perform statistical tests. For the analysis, SPSS version 22 was used to determine the frequencies and percentages, mean differences and paired-samples t-test to rank the most used strategy by all LLMs.

4. FINDING

The study mainly discusses the accuracy and the strategies used in rendering Maribi proverbs. It was found that it is paramount to statistically analyse the outcomes of the LLMs under study by calculating their frequencies, percentages and the mean differences

Table 1: The accuracy of LLMs in the pre-prompt phase

LLMs	totally unacceptable	poor translation	marginally adequate	good translation	very good translation
ChatGPT	41 55.4%	21 28.4%	11 14.9%	0 0.0%	1 1.4%
Gemini	39 52.7%	20 27.0%	12 16.2%	1 1.4%	2 2.7%
DeepSeek	45 60.8%	20 27.0%	7 9.5%	2 2.7%	0 0

In Table 1, the researcher compared accuracy across three prominent LLMs—ChatGPT, Gemini, and DeepSeek —capturing their raw output quality during the pre-prompt phase. Looking at the distribution of the 74 outputs generated by each model. Without the structural guidance of a contextual prompt, each model in this phase struggled profoundly. Almost all their translations clustered at the lowest score of the grading range, revealing a systemic inability to handle Maribi proverbs.

Totally unacceptable and poor translation were the categories that the LLMs clustered around. DeepSeek performed the worst in this pre-prompt phase, with 88% of its total outputs deemed to be totally unacceptable and poor translations. Interestingly, despite completely missing the mark on most attempts, DeepSeek managed to generate "good" translations slightly more often than its peers, though it failed to produce a single "very good" result. Meanwhile, ChatGPT and Gemini flew close behind in overall failure rates.

Gemini managed to generate the most balanced distribution. It recorded the lowest rate of catastrophic failures at 52.7% and actually led the pack in generating acceptable text, pushing just over 4% of its outputs into the good and very good scale. It also captured the highest share of "marginally adequate" translations at 16.2%. ChatGPT, on the other hand, exhibited a rather better performance. This data pointed to the fact that LLMs failed to a great extent to perform better when dealing with the Maribi proverb. Even though Gemini technically edges out the competition by a small margin, a foundational failure rate reached above 50%. Ultimately, this quantitative snapshot serves as concrete proof that robust, intentional prompt engineering isn't just a helpful secondary step—it is an absolute prerequisite to pull these models out of the unacceptable performance score.

Table 2: The accuracy of LLMs in the post-prompt phase

	totally unacceptable	poor translation	marginally adequate	good translation	very good translation
	8	4	32	12	18
ChatGPT	10.8%	5.4%	43.2%	16.2%	24.3%
	11	4	18	18	23
Gemini	14.9%	5.4%	24.3%	24.3%	31.1%
	28	16	25	1	4
DeepSeek	37.8%	21.6%	33.8%	1.4%	5.4%

Table 2 deals with the post-prompting translation accuracy of LLMs, capturing a fundamental shift in their operational efficacy following a targeted instructional prompt. When contrasted with the pre-prompt metrics. The implementation of contextual scaffolding successfully diminished the overwhelming inaccurate translation served in the pre-prompted phase, pulling the majority of the data toward the functional end of the evaluative spectrum. This paradigm shift unequivocally underscores the catalytic role of prompt engineering in optimising generative linguistic capabilities, proving that these systems require active navigational parameters to yield academic-grade material.

Upon closer examination of each LLM separately, Gemini emerged as the most responsive to structural prompting, establishing a clear dominance in high-fidelity output. Evolving significantly from its initial baseline, Gemini managed to push an impressive 55.4% of its translated proverbs into the combined "good" and "very good" brackets. Notably, it secured the highest frequency of top-tier translations, with nearly a third of its total output (31.1%) classified as exceptionally accurate. Its catastrophic failures were radically reduced in this

phase, suggesting an inherent architectural flexibility that allows the model to deeply internalise and execute complex instructional constraints when properly guided.

Table 3: mean comparison of accuracy of pre- and post-prompts of each LLM

		Mean	Std. Deviation	sig
ChatGPT	Pre- Prompt	1.6351	0.83706	0.000
	Post Prompt	3.3784	1.22422	
Gemini	Pre- Prompt	1.7432	0.96601	0.000
	Post Prompt	3.5135	1.37734	
DeepSeek	Pre- Prompt	1.5405	0.77969	0.000
	Post Prompt	2.1486	1.11874	

Table 3 offered a crucial statistical validation of the earlier statistical output by presenting a mean comparison of translation accuracy before and after prompt intervention. Across the board, the data reveal a universally significant shift. For every language model evaluated, the calculated P-values were 0.000. In statistical terms, this revealed that there were statistically significant differences between pre- and post-prompt for all LLMs. Analysing LLMs individually, Gemini and ChatGPT demonstrated remarkable developmental progress. Gemini, which was at the pre-prompt phase at the mean value of 1.7432, jumped to an impressive mean value of 3.5135 after designing the prompt. ChatGPT closely mirrored this evolutionary leap, elevating its initial mean value from 1.6351 to a mean value of 3.3784. This statistical widening indicates that while the instructional scaffolding dramatically raised the overall average quality of the translations, it simultaneously introduced a wider variance in the specific outputs. DeepSeek presents an entirely different analytical narrative. While the system technically achieved a statistically significant improvement—moving from a mean of 1.5405 to 2.1486—the practical utility of this growth is largely insignificant. which fall under “marginally adequate”.

Table 4: Paired Samples Test of means comparison of the accuracy of combated LLMs

		Mean	N	Std. Deviation	Sig.
Pair 1	Pre Prompt	1.6396	74	0.81418	0.000
	post Prompt	3.0135	74	0.93052	

Table 4 presents the predominant findings of the study by presenting a paired-samples t-test that evaluates the combined performance of the language models. By pooling the outputs into a unified framework, this metric allows us to definitively assess the macro-level impact of contextual prompting on generative translation tasks, effectively regardless of individual model biases, to look at the underlying technology as a whole.

Looking closely at the central tendencies of these combined findings, the collective baseline mean was 1.6396 during the pre-prompt. Amazingly, after having the prompt, the mean jumped to 3.0135. This near-doubling of the aggregate score provides compelling, top-down evidence that structured guidance fundamentally alters how these LLMs process and output cross-lingual

data. The data revealed that there was a statistically significant difference, as the p-value was 0.000. This p-value serves as undeniable empirical confirmation that the targeted prompting was the direct, causal mechanism driving the recorded improvements in translation reliability across all LLMs.

Table 5: The strategies used by LLMs in the pre-prompt phase

		Using a proverb of similar meaning and form	Using a proverb of similar meaning but dissimilar form	paraphrasing	Near equivalent or literal translation
ChatGPT	Frq	1	1	11	61
	%	1.4	1.4	14.9	82.4
Gemini	Frq	1	1	9	63
	%	1.4	1.4	12.2	85.1
DeepSeek	Frq	1	1	11	61
	%	1.4	1.4	14.9	82.4

Table 5 shows the choices executed by LLMs during the unprompted translation phase, shedding light on exactly how these models process the text before receiving instructions. The most immediate and striking revelation from this data is the near-absolute homogeneity in their translation. Rather than displaying distinct architectural personalities or diverse problem-solving mechanics, ChatGPT, Gemini, and DeepSeek converged on an almost identical, highly restrictive strategic profile. In fact, ChatGPT and DeepSeek produced perfectly identical frequency distributions, while Gemini deviated by only a fraction of a percent.

By an overwhelming margin, the primary strategy for all three models was "near equivalent or literal translation." Gemini leaned into this mechanical approach the hardest, utilising it for 85.1% of its outputs, while ChatGPT and DeepSeek followed closely at 82.4%. Analytically, this massive overreliance on literal, word-for-word transference directly explains the catastrophic failure rates we previously observed in the baseline accuracy data (see Table 1). When left to their own devices, these LLMs default to superficial lexical decoding. They prioritise structural syntax over semantic depth, effectively ignoring the cultural and contextual nuances that are entirely necessary for translating dialectical proverbs.

Conversely, the models demonstrated a profound inability—or perhaps algorithmic reluctance—to engage in dynamic cultural mapping. The utilisation of culturally equivalent proverbs, whether similar or dissimilar in form, was practically non-existent. Paraphrasing served as a distant, secondary strategy, capturing roughly 12% to 15% of the strategic capacity. This indicates that without a prompt explicitly commanding them to rethink the text's underlying meaning, the models simply could not attempt to reconstruct the message in a more natural target-language format.

Table 6: The strategies used by LLMs in the post-prompt phase

	Using a proverb of similar meaning and form	Using a proverb of similar meaning but dissimilar form	parap hrasin g	Near equivalent or literal translation

ChatGPT	Frq	3	24	31	16
	%	4.1	32.4	41.9	21.6
Gemini	Frq	3	38	17	16
	%	4.1	51.4	23.0	21.6
DeepSeek	Frq	3	2	15	54
	%	4.1	2.7	20.3	73.0

Table 6 displays the strategies employed by the LLMs after having targeted instructional prompts, offering a profound insight into the shifting mechanics of LLMs when subjected to contextual guidance. Contrasted against the severe homogeneity of the baseline phase, this data revealed a radical change in operational behaviours. For the more advanced frameworks, the engineered prompts successfully shattered their initial overreliance on superficial lexical transference, forcing them to adopt significantly more sophisticated, culturally adaptive strategies.

One outstanding remark was that the most compelling evolution was to Gemini, which completely inverted its problem-solving issues and moved away from literal translation in the pre-prompt phase toward dynamic cultural functionality. It elected to utilise proverbs of "similar meaning but dissimilar form" for an overwhelming 51.4% of its translated outputs. This indicates that the model, when properly stimulated, possesses the architectural depth to bypass structural equivalence in favour of profound semantic resonance. ChatGPT underwent a similar process to achieve cultural translation of proverbs to a degree of 32.4%; its primary corrective mechanism was paraphrasing, which dominated 41.9% of its total generations. Consequently, literal translation decreased to a mere 21.6% for both of these leading systems, directly correlating with the surge in higher accuracy scores.

Conversely, DeepSeek's statistical profile in this phase exposed a severe algorithmic rigidity. Despite receiving the same contextual scaffolding- prompt- as its peers, the model largely failed to integrate higher-order strategic thinking. An alarming 73.0% of its outputs remained in the "near equivalent or literal translation" strategy. Its engagement with complex proverbial matching was insignificant, suspended at a statistically insignificant 2.7%. This profound failure to adapt its underlying translation strategy perfectly elucidates why DeepSeek post-prompt accuracy improvements (as shown in Table 2 and Table 3) were so fundamentally constrained compared to the rest of the LLMs.

5. DISCUSSION

This section mainly answered the research objectives.

5.1.Strategies used by LLMs Before and After employ the prompt

Before the introduction of prompts, the cognitive architecture of all three models defaulted to a similar strategy. The baseline data revealed a near-absolute strategic homogeneity: ChatGPT, Gemini, and DeepSeek relied on "near equivalent or literal translation" for over 82% of the Maribi proverbs; this was also noticed by Shehab and Badah (2026) ,who observed that while the correct translation of proverbs is to achieve functional and pragmatic equivalence, LLMs

prioritised formal, word-for -word correspondence. In this phase, the models treated Maribi proverbs as a simple word-for-word equation, completely avoiding the deep cultural mapping required for idiomatic text. Following the integration of targeted prompts, the models were forced into semantic restructuring. The prompts successfully broke the literal translation bias for the more advanced strategies. Gemini aggressively turned, deploying culturally equivalent English proverbs (dissimilar in form) for 51.4% of its outputs. ChatGPT adapted by shifting heavily toward paraphrasing, utilising it as a primary strategy in 41.9% of cases. aligns with Zhang et al. (2023), who added that using parallel prompt examples constructed from monolingual data via zero-shot prompting could improve translation. DeepSeek, however, displayed severe algorithmic rigidity, failing to internalise the new parameters and remaining stuck on literal translation for 73% of its generated text.

5.2.The impact of prompt on the accuracy

This strategic restructuring directly catalysed a massive, statistically validated leap in translational fidelity. During the baseline phase, the models were paralysed by the complexities of the Maribi dialect, clustering heavily in the "totally unacceptable translation" and "poor translation" categories with failure scores around 80% to 88%. These high functional scores go hand in hand with the empirical results of Abdelaziz et al. (2024), who found that training and feeding AI tools made a notable improvement in AI scores, underscoring the value of diverse training corpora in the realm of machine translation. The implementation of the prompts effectively rescued the output from this catastrophic baseline. As was observed in the paired samples t-test, each model surged from low scores to high scores. The finding proves that prompt engineering is not just a supplementary optimisation tool. It is the direct, causal mechanism necessary to force language models out of superficial lexical decoding and into functional, academic-grade translation.

5.3.Evaluating the most benefited LLMs

Gemini was found to be the LLM that benefited most from the prompt in both accuracy and strategy, as Gemini achieved the highest overall accuracy, securing a leading mean score of 3.5135. It also achieved the highest evaluative score, pushing 31.1% of its total translations into the "very good" category, compared to ChatGPT's 24.3% and Deepseek's 5.4%. Gemini stands out as the framework that benefited the most profoundly from the provided prompt. While ChatGPT also successfully doubled its mean accuracy scores as it rose from 1.6351 to 3.3784, this was validated by Beidas et al. (2025), who found that GPT models responded better to prompting and translation to Levant countries compared to Gemini. Gemini's underlying architecture allowed it to execute the most complex instructional constraints. It used the prompt not just to clean up its syntax, but to cope with the dynamic cultural equivalence. Conversely, DeepSeek was found not to benefit greatly from the prompt; despite receiving the same prompt, its underlying structural deficits prevented it from effectively utilising the prompt, despite there was significant differences between the scores before and after the prompt but comparing it to the other two LLMs, it looks weak performance.

6. CONCLUSION

The current study revealed that generative LLMs possess a systemic, foundational deficiency when processing highly localised cultural expressions, specifically Maribi proverbs. These

LLMs failed to generate accurate translations. However, the empirical evidence demonstrates that prompt engineering serves as a transformative cognitive intervention rather than a mere linguistic filter. This targeted scaffolding effectively forces these LLMs to generate better translations. Comparing these three LLMs Gemini was found to be the mostly benefited from the other two despite its weak performance before implying the prompts followed closely by ChatGPT, lastly was that of DeepSeek. regarding to the strategies employed, before the implementation of the prompt also all the LLMs employed literal or near equivalent strategy, but after employing the prompt, the strategy used were changed. Gemini was found to opted to used using a proverb of similar meaning and form or using a proverb of similar meaning but dissimilar form. Conversely, the quiet performance of DeepSeek under identical conditions highlights a critical technological caveat: instructional prompts cannot override fundamental algorithmic limitations. Ultimately, the finding confirmed that prompt engineering is not a supplementary optimisation tactic, but an indispensable prerequisite for extracting academically viable, cross-cultural translations from artificial intelligence.

6.1.Pedagogical Implications

For translation training, applied linguistics, and curriculum development, these findings necessitate an immediate paradigm shift in how AI is conceptualised within the classroom. So educators must move beyond treating LLMs as automated dictionaries. Translation programs need to explicitly integrate prompt engineering into their syllabi, training students to design the specific contextual guardrails required to guide models toward cultural equivalence and nuanced paraphrasing. The profound divergence in capability between Gemini, ChatGPT, and DeepSeek underscores the necessity of critical AI literacy. Students must be taught to evaluate the architectural flexibility of different generative platforms, recognising that models are not uniformly capable and that selecting the right framework is just as critical as writing the right prompt.

6.2.Recommendations for Future Studies

While this analysis successfully maps the behavioural mechanics of LLMs translating Maribi proverbs, the rapid evolution of generative AI opens several necessary avenues for subsequent research. Future investigations should broaden the linguistic parameters to include other highly localised and culturally dense Arabic dialects, such as Hodeidi, Tehami, or Adeni variations. This would determine whether the observed algorithmic rigidities are unique to the Maribi dialect or indicative of a broader systemic failure in cross-linguistic regional processing.

REFERENCES

Abdelaziz, A. A. A., Elneima, A. H., & Darwish, K. (2024). LLM-based MT data creation: Dialectal to MSA translation shared task. In *Proceedings of the 6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT) with Shared Tasks on Arabic LLMs Hallucination and Dialect to MSA Machine Translation@ LREC-COLING 2024* (pp. 112-116). https://aclanthology.org/2024.osact-1.14.pdf

Ahmed, M. A. (2025). From text to understanding the inner text: LLMs and translation accuracy and fluency. *International Journal of Language and Literary Studies*, *7*(2), 139-156. http://doi.org/10.36892/ijlls.v7i2.207

Ahmed, M. F. (2010). An Intertextual Approach to Arabic-English Translation. Basic *Education Education Researchers Journal*, 10(1), 505–527.

Al Rousan, R. M., & Shatnawi, H. (2023). The Translation of Women-Related Proverbial Expressions from Jordanian Arabic into English: An Error Analysis. *Jordan Journal of Modern Languages & Literatures*, *15*(2), 447-472. https://doi.org/10.47012/jjmll.15.2.4

Alfaleh, B. A. (2017). Translation Quality Assessment of Proverbs from English into Arabic: The Case Study of One Thousand and One English Proverbs Translated into Arabic. (M.A. Thesis). Retrieved from *Arab World English Journal* (ID Number: 250. May 2020, 1- 56.DOI: https://dx.doi.org/10.24093/awej/th.250

Alharbi, A. S. (2023). Translating Saudi Najdi Dialect Proverbs into English: Challenges and Strategies for Preserving Cultural Meaning. *International Journal of Linguistics, Literature and Translation, 6*(12), 213–227. https://doi.org/10.32996/ijllt.2023.6.12.25

Al-Hussami, A. A., & al-Hiba, M. (2023). Gender Discrimination against Women in Yemeni Proverbs. *Mahrah University Journal of Humanities*, *4*(2), 461–482. https://doi.org/10.71311/.v4i2.127

Almjlad, S. A. (2023). How the Saudi EFL Students at undergraduate level translate English proverbs into Arabic: Problems, strategies and solutions. *AWEJ for Translation & Literary Studies*, *7*(4).

Al-Saidi, A. H. S. (2021). Pragmatic analysis of the translation of English culture-specific proverbs into Arabic. *Journal of the College of Languages (JCL)*, (43), 30-63. https://doi.org/10.36586/jcl.2.2021.0.43.0030

Al-Saqqaf, A. H. (2023). Plant and Animal Terms in Ḥaḍramī Arabic Idiomatic Expressions, Proverbs and Chants. *Bloomsbury Advances in Ecolinguistics Series Editors*, 167.

Al-Wasy, B. Q., & Mohammed, O. S. M. (2024). Strategies of Translating Euphemistic Expressions from Arabic into English: A Comparative Study of Artificial Intelligence Models with Human Translation. *Humanities and Educational Sciences Journal*, (40), 826–855. https://doi.org/10.55074/hesj.vi40.1121

Ataman, D., Birch, A., Habash, N., Federico, M., Koehn, P., & Cho, K. (2025). Machine translation in the era of large language models: a survey of historical and emerging problems. *Information*, *16*(9), 723. https://doi.org/10.3390/info16090723

Bajubair, A., & Abdulla, N. A. (2022). The Translatability of Hudhrami Women Proverbs into English among MA Translation Students at Al-Rayan University. *Al-Andalus journal for Humanities & Social Sciences*, *9*(55), 220.

Baker, M. (1992). *In other words: A coursebook on translation*. London and New York: Routledge.

Baker, M. (2018). *In other words: A coursebook on translation*. Routledge. Third Edition. https://doi.org/10.4324/9781315619187

Beidas, A., Mohi, K., Ghaddar, F., Ahmad, I., & Abed, S. E. (2025). Cross-dialectal Arabic translation: comparative analysis on large language models. *Frontiers in Artificial Intelligence*, *8*, 1661789. https://doi.org/10.3389/frai.2025.1661789

Budimir, B. (2025). The challenge of translating culture-specific items: Evaluating MT and LLMs compared to human translators. In *Proceedings of Machine Translation Summit XX: Volume 1* (pp. 455-467). https://aclanthology.org/2025.mtsummit-1.36/

Dickins, J., Hervey, S., & Higgins, I. (2016). *Thinking Arabic translation: A course in translation method: Arabic to English*. Routledge.

Ertugruloglu, O., Gülcan, N. Y., & Abidoğlu, Ü. P. (2024). The impact of culture on the education system. *International e-Journal of Educational Studies*, *8*(16), 131-141. http://doi.org/10.31458/iejes.1436091

Farghal, M., & Saeed, A. T. (2022). Proverbial use of fauna in Arabia: A theoretical, thematic and translational account. *Journal of Southwest Jiaotong University*, *57*(6), 453-465.

Gain, B., Bandyopadhayay, D., Ekbal, A., & Singh, T. N. (2026). Bridging the linguistic divide: a survey on leveraging large language models for machine translation. *Language Resources and Evaluation*, *60*(2), 40. https://doi.org/10.1007/s10579-026-09919-7

Hazaea, A. (2017). Human values in intercultural communication: CDA for discourse of proverbs in Yemen Times. *Idealogy Journal*, *2*(2), 34-52. https://doi.org/10.24191/idealogy.v2i2.24

Jabak, O. (2022). Proposed Taxonomy of Strategies for Translating English Proverbs into Arabic. *Journal of Translation and Language Studies*, *3*(2), 37-49. https://doi.org/10.48185/jtls.v3i2.507

Jiang, Z., Lv, Q., Zhang, Z., & Lei, L. (2024). Convergences and divergences between automatic assessment and human evaluation: Insights from comparing ChatGPT-generated translation and neural machine translation. *arXiv preprint arXiv:2401.05176*.

Jibreel, I. (2023). Online machine translation efficiency in translating fixed expressions between English and Arabic (proverbs as a case-in-point). *Theory and Practice in Language Studies*, *13*(5), 1148-1158. https://doi.org/10.17507/tpls.1305.07

Kamal, N. (2022). Verbal Art as a Route into Local Knowledge in Yemen: The Work of ʿAbd Allāh al-Baraddūnī on ʿAlī bin Zāyid. *Narrative Culture*, *9*(1).

Kukreja, S., Kumar, T., Purohit, A., Dasgupta, A., & Guha, D. (2024, January). A literature survey on open source large language models. In *Proceedings of the 2024 7th International Conference on Computers in Management and Business* (pp. 133-143). https://doi.org/10.1145/3647782.3647803

Kumar, P. (2024). Large language models (LLMs): survey, technical frameworks, and future challenges: P. Kumar. *Artificial Intelligence Review*, *57*(10), 260. https://doi.org/10.1007/s10462-024-10888-y

Lau, O. E., Amini, M., Siau, C. S., & Tan, D. A. L. (2022). Translation of Chinese Culture Specific Items to English in the Film Crouching Tiger, Hidden Dragon. AJELP: *Asian Journal of English Language and Pedagogy*, *10*(2), 68-81.

Mahdy, O. S. M.., Samad, S. S., & Mahdi, H. S. (2020). The attitudes of professional translators and translation students towards computer-assisted translation tools in Yemen. *Journal of Language and Linguistic Studies,16*(2), 1084-1095.

Mohammed, T. M. Q., & Ho-Abdullah, I. (2021). Universality and language specificity: Evidence from Arab and English proverbs. *International Journal of Comparative Literature & Translation Studies*, *9*(1), 24.

Nida, E. A. (1964). *Toward a science of translating: With special reference to principles and procedures involved in Bible translating*. Brill Archive.

Noman, A., Almansoob, N. ., Mohammed, O. S. M. ., & Alrefaee, Y. . (2025). AI-Driven and Large Language Models-Based Translation of Arabic News Texts into English: A Comparative Evaluation. *Journal of Translation and Language Studies*, *6*(3), 40–54. https://doi.org/10.48185/jtls.v6i3.1945

Qasim, R. A. A. G. (2026). Strategies Used for Translating Yemeni Colloquial Proverbs from Arabic into English and Their Effects on Translation. *Global Journal of English & Translation Studies*, *1*(15), 73-85.

Qian, M., & Kong, C. (2024, May). Enabling human-centered machine translation using concept-based large language model prompting and translation memory. In *International Conference on Human-Computer Interaction* (pp. 118-134). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-60615-1_8

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of things and cyber-physical systems*, *3*, 121-154. https://doi.org/10.1016/j.iotcps.2023.04.003

Saideen, B., Thalji, A., Lababaneh, A., Tartory, R., Haider, A. S., & Obeidat, M. (2024). Netflix English subtitling of the Jordanian movie 'The Alleys': Challenges and strategies. *Research journal in advanced humanities*, *5*(1), 157-177.

Shehab, E. (2026). Decoding the unspoken: Pragmatic and cultural challenges in translating Arabic proverbs into English. In *FORUM* (Vol. 24, No. 1, pp. 87-108). Amsterdam/Philadelphia: John Benjamins Publishing Company.

Shehab, E., & Badah, A. M. (2026). Translating conversational implicatures: a comparative study of human and artificial intelligence performance in Arabic–English literary translation. *Translation Review*, *124*(1), 38-54. https://doi.org/10.1080/07374836.2026.2617670

Shhaiber, R., & Haider, A. S. (2023). Strategies of subtitling Egyptian idiomatic expressions and proverbs into English. *Language Value*, *16*(2), 60-99. https://doi.org/10.6035/languagev.7015

Villers, D. (2022). What makes a good proverb? On the birth and propagation of proverbs. *Lexis. Journal in English Lexicology*, (19). https://doi.org/10.4000/lexis.6383

Wang, H., & Wu, Y. (2026). Large Language Model-Driven Intelligent Translation: How New Technologies Are Reshaping Translation Practices?. In AI Through LLMs: Transforming Translation Practice and Teaching (pp. 3-29). Singapore: Springer Nature Singapore.

Wang, Y., Zhang, J., Shi, T., Deng, D., Tian, Y., & Matsumoto, T. (2024). Recent advances in interactive machine translation with large language models. *Ieee Access*, *12*, 179353-179382. https://ieeexplore.ieee.org/abstract/document/10737082

Zhang, B., Haddow, B., & Birch, A. (2023, July). Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning* (pp. 41092-41110). PMLR. https://proceedings.mlr.press/v202/zhang23m.html

AUTHOR'S BIO

Othman Saleh Mahdy Mohammed ***is an Assistant Professor. He has published several articles in international journals, and he has also worked as an editor and reviewer for many journals indexed in Scopus and ISI. Scopus ID: 57219415483***