Can LLMs Translate Cultural Nuance in Dialects? A Case Study on Lebanese Arabic

Silvana Yakhni, Ali Chehab
Electrical and Computer Engineering
American University of Beirut
syy06@mail.aub.edu, chehab@aub.edu.lb

Abstract

Machine Translation (MT) of Arabic-script languages presents unique challenges due to their vast linguistic diversity and lack of standardization. This paper focuses on the Lebanese dialect, investigating the effectiveness of Large Language Models (LLMs) in handling culturally-aware translations. We identify critical limitations in existing Lebanese-English parallel datasets, particularly their non-native nature and lack of cultural context. To address these gaps, we introduce a new culturally-rich dataset derived from the *Language Wave (LW)* podcast. We evaluate the performance of LLMs: *Jais*, *AceGPT*, *Cohere*, and *GPT-4* models against Neural Machine Translation (NMT) systems: *NLLB-200*, and *Google Translate*. Our findings reveal that while both architectures perform similarly on non-native datasets, LLMs demonstrate superior capabilities in preserving cultural nuances when handling authentic Lebanese content. Additionally, we validate *xCOMET* as a reliable metric for evaluating the quality of Arabic dialect translation, showing a strong correlation with human judgment. This work contributes to the growing field of Culturally-Aware Machine Translation and highlights the importance of authentic, culturally representative datasets in advancing low-resource translation systems.

1 Introduction

The Arabic script, known for its use in writing Modern Standard Arabic (MSA), is used by hundreds of millions of people worldwide across a diverse range of languages, including Arabic dialects, Abjad, and Ajami languages. Arabic-script languages share several key characteristics, including a rich cultural context, idiomatic expressions, and frequent use of religious and poetic references. These features make translation particularly challenging, as they require not only linguistic accuracy but also cultural sensitivity. This paper focuses on Lebanese Arabic, a prominent dialect spoken in the Levant region, that exemplifies the script complexities, with its unique cultural expressions and idioms.

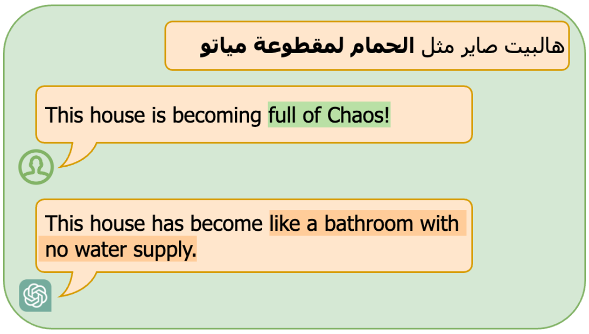

هالبيت صاير مثل **الحمام لمقطوعة مياتو**
This house is becoming full of Chaos!
This house has become like a bathroom with no water supply.

Figure 1: Example of the translation of the Lebanese idiom (الحمام لمقطوعة مياتو) by a human translator compared to GPT-4o

However, the predominantly spoken nature of dialects, coupled with their lack of standardized spelling and grammar, presents a significant challenge for Machine Translation (MT) due to the scarcity of culturally representative datasets needed to develop effective translation models. The few available parallel Lebanese/English data suffer from many limitations, including the predominance of foreign source languages in existing corpora (Krubiński et al., 2023) (Bouamor et al., 2018) (team et al., 2022), which may not accurately capture the nuances of the Lebanese culture.

Recently, Decoder-only Large Language Models (LLMs) such as chatGPT[1], Claude[2], and LLaMA (Touvron et al., 2023) have demonstrated notable success across various

[1] https://chatgpt.com/
[2] claude.ai

Processings of the 1st Workshop on NLP for Languages Using Arabic Script (AbjadNLP), pages 114–135
January 19, 2025. ©2025 Association for Computational Linguistics

NLP tasks, including translation, particularly for widely used languages (Jiao et al., 2023)(Lyu et al., 2023). Recent research has tackled Culturally-Aware Machine Translation (CAMT) (Yao et al., 2024) with LLMs and showed that they exhibit superior capabilities compared to traditional neural MT systems in translating cultural content.

In Arabic NLP, little effort was made to benchmark the performance of LLMs in translating Arabic dialects. However, these efforts fell short of assessing the full spectrum of Arabic-focused LLMs (Kadaoui et al., 2023)(Alam et al., 2024). Furthermore, Existing Arabic dialect evaluation benchmarks such as LAraBench (Abdelali et al., 2023), SADID (Abid, 2020) and AraDICE (Mousi et al., 2024) rely primarily on translated English content, rather than authentic dialectal resources. This limitation extends beyond isolated cultural elements to the entire linguistic system, including culturally embedded grammar, vocabulary, and idioms. Figure 1 shows a failed attempt of *GPT-4o* to translate the cultural Lebanese idiom ”el-hamem el-maa’toua’a maytu” (الحمام لمقطوعة مياتو) , which means **”It’s Chaos”**. GPT-4o instead literally translates it to ”a bathroom with no water supply”. The field’s dependence on translated data underscores the urgent need for developing authentic, culturally-aware datasets that capture the true complexity of Arabic dialectal variations. Appendix B provides a more comprehensive overview of previous research in this domain.

Moreover, the evaluations of translation tasks for Arabic dialects depend mainly on statistical metrics like the BLEU score, despite substantial evidence showing its limitations in evaluating fluency and meaning compared to neural metrics such as xCOMET(Kocmi et al., 2024)(Lee et al., 2023).

More specifically, in this work, we aim to answer the following questions:

1. Do existing Lebanese-English datasets accurately reflect translation quality, given their English origins and limited Lebanese cultural context?

2. Do LLMs and encoder-decoder models perform equally across all datasets, or do they struggle with culturally rich datasets?

3. Which performs better in translating Arabic dialects: LLMs or translation NMT models?

To this end, we review the few existing parallel Lebanese/English datasets and critically assess their shortcomings. We then introduce our new curated dataset from the Language Wave (LW) podcast, a collection of culturally rich Lebanese content, and we demonstrate how this dataset effectively addresses the limitations of existing resources by ensuring cultural authenticity, a trait typically absent in datasets derived from non-native sources. Through a comprehensive comparative analysis, we evaluate closed-source Arabic-focused LLMs (Jais (Sengupta et al., 2023), AceGPT (Huang et al., 2024), Cohere [3]) and the API-based model (GPT-4o[4]) against open-source NMT systems (NLLB-200) (team et al., 2022) and commercial translation services (Google Translate), examining their performance on culturally-rich Lebanese content versus English-derived datasets.

Our key findings demonstrate several significant insights:

- A systematic analysis reveals a substantial gap in existing parallel datasets regarding cultural representation.

- While Arabic-focused LLMs and encoder-decoder models exhibit comparable performance on traditional English-origin datasets, LLMs remarkably demonstrate superior performance when processing culturally-aware datasets.

- Open-source Command-R+ rivals GPT-4o in cultural translation, promoting accessible tools.

- We demonstrate the effectiveness of xCOMET as a reliable evaluation tool to assess the translation quality from Arabic dialects to English, with results showing a high correlation with human judgment.

2 Existing Datasets

Open Subtitles(OS) (Krubiński et al., 2023): A large dataset containing 120,600 sentences derived from movie subtitles, available

[3] https://cohere.com/

[4] https://chatgpt.com/

in both MSA and English. Researchers manually translated MSA sentences into Lebanese. Despite its size, it has significant quality issues stemming from using Modern Standard Arabic (MSA) as an intermediary language for translations. In addition, the dataset suffers from cultural misalignment given its translation from Western-centric source material. We refer to this data as OS (Open Subtitles).

MADAR CODA (Bouamor et al., 2018): A corpus containing 2,000 English sentences from the Basic Travel Expression Corpus (BTEC) translated into 26 Arab city dialects, with expanded coverage of 10,000 sentences for major cities, including Beirut. While valuable for dialectal variation, the dataset is limited by its simple sentence structures and its narrow focus on travel-related content. The English-sourced translations also potentially introduce cultural bias, limiting its effectiveness for culturally-aware machine translation applications.

Facebook Low Resource (FLoRes) Corpus (team et al., 2022): A benchmarking dataset containing 3,001 sentences from Wikimedia projects, professionally translated into over 200 languages. While broad in language coverage, the dataset's formal content lacks the informal linguistic features and cultural nuances essential for dialect translation.

Arabic-Dialect/English Parallel Text (Zbib et al., 2012): A substantial corpus developed through collaboration between Raytheon BBN Technologies, LDC, and Sakhr Software, containing 3.5 million tokens of Arabic dialectal content with English translations, focusing on Levantine and Egyptian dialects. While potentially valuable, its restricted access through LDC has limited its research impact, with no comprehensive quality evaluation existing in the literature.

3 Language Wave Dataset

The development of a parallel Lebanese Arabic-English dataset addresses critical gaps in existing translation resources for this dialect. Our comprehensive data collection process focused on creating an authentic, diverse, and professionally translated corpus that effectively captures the nuances of Lebanese Arabic while providing professional English translations. Through careful curation of Lebanese media sources, we prioritized maintaining cultural relevance and linguistic authenticity, ensuring the dataset would serve as a valuable resource for both academic research and practical applications.

We identified the "Language Wave" podcast[5] as an invaluable resource in preserving cultural content. This podcast, with its slogan **"Learn Lebanese Arabic with transcribed podcast: episodes exploring Lebanon and its people"**, offers authentic content that covers various topics and language concepts, designed to enhance Lebanese Arabic skills in active listening, reading, vocabulary, and cultural context knowledge. Through collaboration with the "Language Wave" podcast, we developed a comprehensive dataset encompassing 95 episodes, which resulted in 2,947 Lebanese sentences professionally translated into English. The podcast's colloquial style effectively mirrors everyday Lebanese Arabic conversations and mimics authentic, colloquial Lebanese Arabic. **We refer to our Language Wave dataset as "LW".**

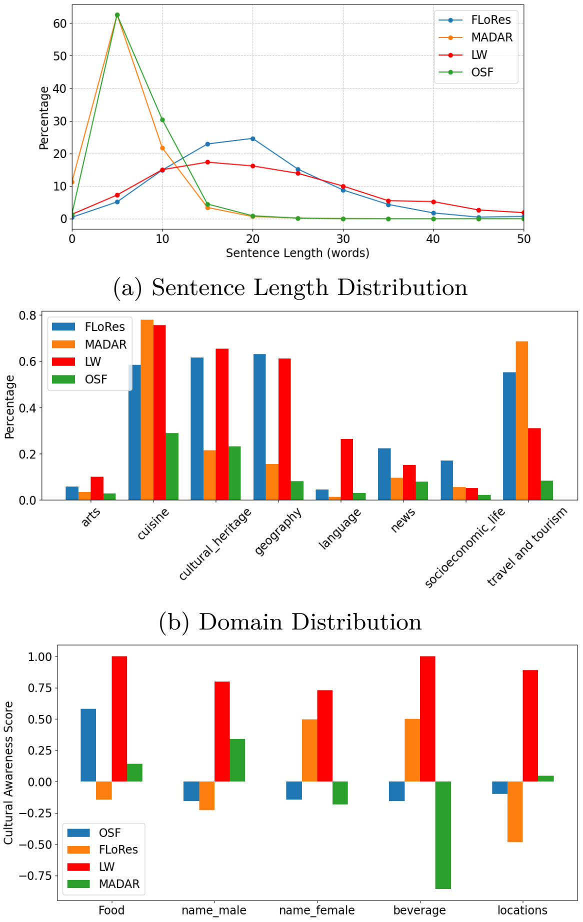
4 Linguistic Analysis

LW dataset exhibits several distinguishing characteristics when compared to MADAR, FLoRes, and OS. The most significant attribute is data authenticity among others. While the aforementioned datasets are translated from foreign sources, LW is uniquely crafted by professional translators, ensuring a high degree of linguistic fidelity. To highlight the distinctive features of the LW dataset, we conduct comprehensive analyses, the results of which are presented in Figure 2.

1. **Sentence Length Distribution:** Analysis of sentence length distribution reveals that LW exhibits a more balanced spread across various lengths, indicating a more natural and varied language usage.

2. **Domain Distribution:** We compiled a comprehensive lexicon encompassing 8 prominent domains: arts, cuisine, cultural heritage, geography, language, news, socioeconomic life, and travel and tourism. For

[5] https://languagewave.com/



(a) Sentence Length Distribution

(b) Domain Distribution

(c) Cultural Awareness

Figure 2: Comparative Data Analysis between MADAR, FloRes, LW and OS Datasets based on three criteria: Domain Distribution, Cultural Awareness, and Sentence Length distribution.

each dataset, we calculated the frequency of domain words. LW demonstrates robust representation in critical categories such as cuisine, cultural heritage, and geography. Notably, MADAR exhibits a bias towards the travel and tourism domain, while OS shows the least richness across all domains, potentially due to its nature as movie subtitles. This distribution underscores the rich diversity inherent in our dataset.

3. **Cultural Awareness:** To quantify this crucial characteristic, we employed a Cultural Awareness metric, inspired by the work in (Naous et al., 2023) to assess the cultural awareness of LLMs. We selected five domains $D = d_1, ..., d_5$ where d_1 = ”Food”, d_2 = ”male names”, d_3 = ”female names”, d_4 = ”beverages”, and d_5 = ”locations”. For each domain $d_i \in D$ and dataset X, we calculated the frequency of Arab terms (f_A) and Western terms (f_W) through exact string matching. The Cultural Awareness Score (CAS) for each domain is defined as:

$$CAS(d_i) = \frac{f_A(d_i) - f_W(d_i)}{f_A(d_i) + f_W(d_i)} \in [-1, 1] \quad (1)$$

where $f_A(d_i)$ and $f_W(d_i)$ represent the frequency of Arab and Western terms respectively in domain d_i. The results demonstrate that LW consistently achieves high positive CAS values for all categories, particularly excelling in name recognition (both male and female) and locations. This exceptional performance distinctly sets LW apart, indicating its superior ability to capture nuanced cultural context.

CAS as a Cultural Benchmark : The Cultural Awareness Score (CAS) provides an initial benchmark for quantifying cultural representation in linguistic datasets, while simultaneously acknowledging the inherent complexities of cultural linguistic analysis. Although the metric employs a binary classification of Arab and Western terms, its primary value lies in establishing a structured methodology for examining cultural nuances in low-resource datasets. To enhance the metric's adaptability, a great approach is to add on the Arab terms we have by collaborating with linguistic experts who can provide comprehensive compilations of region-specific expressions, idioms, and cultural references that might otherwise be overlooked in standard linguistic analyses. This approach allows for potential adaptation to other language contexts by leveraging expert knowledge in cultural linguistics, and translation studies.

5 Quantitative Analysis

The core question guiding our analysis is the following: **How proficient are translation models in producing translations that preserve cultural nuances and context?**

To address this question, we leverage our Lebanese culturally-aware dataset, Language Wave (LW), to assess the translation performance of both decoder-only and encoder-

	Non-native			Culturally-Aware
	FLoRes	OS	MADAR	LW
NLLB-3.1B	0.88463969	0.87022187	0.86595088	0.63533914
NLLB-moe-54B	0.89736591	0.92584903	0.88523401	0.65198355
Google-Translate	0.92879412	0.92916107	0.88343378	0.66453003
Jais-13B	0.84704429	0.84447611	0.88901745	0.70714775
Jais-adapted-70B	0.89354683	0.92794033	0.91857263	0.75139506
AceGPT-7B	0.79234672	0.81954603	0.82858921	0.66264395
AceGPT-70B	0.85027576	0.86379206	0.87928573	0.75365206
Aya32-8B	0.87830721	0.90754499	0.87050557	0.68780926
Aya-expanse-32B	0.90185826	0.91843578	0.89275793	0.75132963
Command-R+	0.92475206	0.92651753	0.92847502	0.80957264
GPT-4o	0.93451795	0.93367150	0.92774067	0.79337348

Table 1: Comparative assessment of translation quality across encoder-decoder architectures (NLLB, GoogleTranslate) and Large Language Models (Jais, AceGPT, Cohere, GPT-4o). The analysis spans three established non-native benchmarks (FLoRes, MADAR, OS) and our culturally-aware LW dataset, measuring xCOMET scores between reference and generated translations.

decoder models. Additionally, we compare their performance when translating three non-native datasets — FLoRes, MADAR, and OS. For evaluation, we conducted a thorough correlation analysis in section 7. Our results show that xCOMET shows the highest correlation with human judgment. Hence, we adopt in this work xCOMET-10.7B as our evaluation metric.

MT systems in Comparison: We evaluate the following MT systems:

- NMTs: We evaluate the state-of-the-art multilingual NLLB models: *NLLB-3.1B* and *NLLB-moe-54B*. We also use the *Google-Translate* engine in our comparison.
- LLMs: We examine the following Arabic-focused open-source models: *Jais-13B*, *Jais-adapted-70B*, *AceGPT-7B*, *AceGPT-70B*, in addition to the multilingual open-source Cohere models: *Aya23-8B*, *Aya-expanse-35B* and *Command-R+-104B*. Finally, we used the closed API-based *GPT-4o* model. More details about the models are available in Appendix A.

Experimental Results: For decoder-only models, we prompted the model as follows: *"You are a professional translator, translate the following sentence from Lebanese to English: Input: {sentence}"*. In this study, we focused solely on zero-shot prompting for LLMs and used encoder-decoder models without fine-tuning. This approach was chosen to evaluate the innate capability of these models to comprehend and translate culturally rich and nuanced content without relying on task-specific training.

The second question we aim to answer in this analysis is: **How do the performance of LLMs and encoder-decoder models compare, when handling culturally-aware content?** Our analysis reveals intriguing patterns: for content derived from Western cultures (MADAR, FLoRes, OS), both architectures demonstrate comparable performance, with encoder-decoder models like NLLB-moe-54B and Google-Translate achieving scores that occasionally surpass decoder-only models like Jais-adapted-50B, Command-R+. However, a notable divergence emerges when handling culturally rich Lebanese content. LLMs consistently outperform NLLB and Google-Translate on culturally-aware datasets. While Jais-adapted-70B and Command-R+ maintain scores in range (0.75-0.8) on LW's cultural examples, encoder-decoder models' performance drops significantly to a range of around 0.65. These findings suggest that the architectural advantages of LLMs may be particularly valuable for preserving cultural nuances in translation, though further research is needed to fully understand this phenomenon. In addition, our analysis reveals a clear correlation between LLM size and translation quality, as measured by xCOMET scores. Larger models like Jais-adapted-70B, AceGPT-70B, and Command-R+ consistently outperformed their smaller counterparts. Notably, the 104B Command-R+ achieved comparable results to GPT-4, even exceeding it on the LW dataset. These findings suggest promising opportunities for developing accessible, high-quality cultural translation tools.

6 Qualitative Analysis

To complement our quantitative findings, we conducted a qualitative analysis focusing on four distinct aspects of Lebanese-English cultural translation: **1) cultural understanding, 2) linguistic complexity, 3) idiomatic language, and 4) Ambiguity in translation**. We tested the translation of Lebanese expressions on four different models. For encoder-decoder models, we chose *Google-Translate.* For LLMs, we tested the closed *GPT-4o* model, the multilingual *Command-R+-104B*, and the Arabic-focused *Jais-adapted-70B.* Some of these examples are highlighted in figures 3-6 in Appendix D.

Cultural Understanding: Our initial analysis examined terms that represent various aspects of Lebanese culture, including religious references, social customs, and traditional practices. A notable example, shown in Figure 3, involves social custom phrases such as "katb el-kteb" (كتب الكتّاب), denoting the formal marriage contract announcement, "el-mokaddam" (المقدم), referring to the bride's initial dowry, and "el-moa'khar" (المأخر), indicating the deferred dowry allocated to the bride in case of divorce. While Google Translate employed a literal translation approach that failed to convey cultural significance, LLMs exhibited enhanced comprehension of cultural nuances, with Command-R+ demonstrating exceptional translation accuracy that surpassed even GPT-4o. Furthermore, we tested the models' cultural understanding on the Lebanese term "el-sett el-marje'youniye" (الست المرجعيونية), which translates to "the lady from Marje'youn"- "el-marje'youniye" (المرجعيونية) is an adjective derived from the Lebanese village noun "Marje'youn" (مرجعيون). We notice that Command-R+ was able to convey this meaning in its translation, while also preserving the tone of respect by translating (الست) to "lady" rather than "woman".

Linguistic Complexity: To assess linguistic complexity, we extracted challenging sentences from a Lebanese vocabulary textbook, focusing on grammatical structures and vocabulary unique to the Lebanese dialect. This analysis revealed that while models could effectively handle basic dialectal variations, they encountered difficulties with unique Lebanese vocabulary. A particularly illustrative challenge emerged in the translation of "Lebanized" verbs (non-Semitic verbs that have been morphologically adapted to Lebanese linguistic patterns). Figure 4 presents the example of such a verb- "mdapras" (مدپرس), which means "got depressed." Furthermore, Lebanese Arabic is characterized by distinctive terms that often carry subtle contextual implications. As demonstrated in Figure 4, the term "anja'" (أنجا) emphasizes a narrow escape or marginal success, typically carrying undertones of fortunate timing. While Google Translate failed to convey the meaning accurately, LLMs performed significantly better, with Command-R+ particularly successful in capturing the subtle undertones, translating (أنجا) as "barely managed" rather than "managed." Similarly, the Lebanese term "yestefil" (يصطفل) conveys indifference or detachment regarding another person's situation or decision, often implying personal responsibility for consequences and carrying a tone of irritation. While Google Translate struggled significantly with this term, LLMs demonstrated superior comprehension. Notably, while GPT-4 incorrectly translated this term as "suit yourself," Jais and Command-R+ provided more accurate translations with "Let him be."

Idiomatic Language: Our third analysis examined the use of Lebanese idioms, with particular attention to everyday expressions. A representative example shown in Figure 5 is "ana bi wadi w inti bi wadi" (أنا بوادي وإنتي بوادي), literally meaning "I am in a valley and you are in a valley". This phrase is used to indicate a significant disconnect between two parties' perspectives and is often translated literally by Google Translate, resulting in the loss of its cultural significance. Similarly, the idiomatic expression "hases hali metl la'trach bzaffe" (حاسس حالي متل الأطرش بالزفة), literally translates to "I feel like a deaf person in a wedding ceremony", but usually means "I feel out of place". Note that LLMs are usually able to describe situations where an idiom is used, which opens horizons for exploring different prompting techniques that can guide LLMs to translate culturally-aware expressions.

Ambuiguity: Translation in Arabic and Abjad scripts can be ambiguous due to the absence of diacritics, which leaves words open to multiple interpretations based on context. Additionally, using adverbs connected to verbs can alter meaning subtly, making it difficult for machine translation systems to capture their intended use. Examples of ambiguous translations are shown in Figure 6. The Arabic word (كتبت), can be transcribed based on diacritics as "katabet" or "katabit", meaning "I wrote" or "she wrote", depending on the context. Another example is the reference to an adverb; the expression "el-walad wa'aa' a'n lkersi fankasaret e'jru" (الولد وقع عن لكرسي فنكسرت اجرو), translates to "The boy fell from the chair and he broke **his/its** leg". Despite strategic attempts to disambiguate these terms and provide contextual clarity, both Google Translate and LLMs failed to provide correct translations.

Our comparative analysis of Lebanese-English translation models reveals a clear hierarchy in translation capabilities, with Command-R+ and GPT-4o consistently outperforming other models across cultural, linguistic, and idiomatic dimensions, while traditional encoder-decoder models like Google Translate showed significant limitations and often fail to capture cultural significance. Despite the clear advantage of LLMs, they still struggle in many scenarios, especially in idiomatic and ambiguous settings.

7 Cultural Translation Landscapes

Our methodological approach for Lebanese dialect translation provides a framework for addressing challenges in low-resource languages, especially those using Arabic scripts, given the common linguistic challenges they face, including diacritization, lexical ambiguity, and preserving culturally embedded expressions.(Ishaku et al., 2020).

Another common challenge is the lack of carefully curated, culturally-rich datasets. A few notable examples include the Curras+Baladi dataset(Haff et al., 2022), which focuses on translating authentic songs and blog posts for the Levantine dialect. Efforts were also made to collect such datasets in Egyptian (Al-Sabbagh, 2023). Furthermore, the Boston University research project on Ajami Literacy, supported by the National Endowment for the Humanities, has made significant strides by digitizing manuscripts in four West African languages (Hausa, Mandinka, Fula, and Wolof), providing transcriptions, translations, and multimedia resources (Ngom et al., 2023). Despite these efforts, existing linguistic resources remain insufficient to comprehensively address the complexities of translating Arabic-script languages.

Building upon our analysis of Lebanese dialect translation, this study made an additional effort to explore some of the linguistic commonalities across other Arabic-script languages, with a specific focus on Hausa and Wolof Ajami languages. Our analysis concentrates on the nuanced translation of idiomatic expressions, culturally specific terminology, and religious lexicons. Comparative translation examples for both Hausa and Wolof from GPT-4o and Google Translate, are detailed in Appendix D and illustrated in Figures 7-10, providing a comprehensive examination of challenges inherent in these culturally-rich low-resource languages. All examples are taken from resources in (Ngom et al., 2023). Similarly to Lebanese, preliminary findings on Hausa and Wolof reveal that LLMs demonstrate notable limitations in accurately interpreting cultural expressions, though they exhibit marginally superior performance compared to Google Translate. These results underscore the critical need for further comprehensive linguistic analysis that moves beyond mere lexical conversion to a more profound understanding of cultural meaning-making processes.

8 Metric Correlation Analysis

Machine translation evaluation relies on numerous established metrics, each with its own strengths and methodologies. While learned neural metrics like COMET(Rei et al., 2022) and BERTScore (Zhang et al., 2019) have demonstrated superior correlation with human judgment compared to traditional metrics like BLEU (Kocmi et al., 2024)(Lee et al., 2023), the latter is still widely used in Arabic NLP. To evaluate the effectiveness of different automatic metrics for Lebanese dialect to English translation, we conducted a correlation analysis with human judgment. The experiment was designed to balance between rigor and resource constraints.

Metrics to evaluate: BLEU(Papineni et al., 2002), BERTScore(Zhang et al., 2019), COMET(XLM-R Large)(Rei et al., 2022), and xCOMET-10.7B(Guerreiro et al., 2023). More details are provided in Appendix C.1.

Dataset: We conducted a human evaluation study using 150 sampled sentence pairs from our Lebanese Arabic (LW) dataset. The sample was strategically selected to ensure authentic Lebanese content and balanced representation across various linguistic phenomena and complex grammatical structures, as well as diverse domain topics. For our evaluation, we chose to focus on translations generated by the Aya23-8B model. This decision was motivated by our aim to obtain meaningful human ratings across the full spectrum of translation quality (good, acceptable, and poor). While models like GPT-4o[6] and larger architectures such as Command-R+[7] typically produce high-quality translations, and NLLB-1.5B(team et al., 2022) often contains numerous errors, Aya23-8B generates translations with sufficient variation in quality to facilitate nuanced human evaluation.

Human Annotation Guidelines: The translations of the 150 sentences were subsequently subjected to human assessment to evaluate their quality. Three bilingual annotators, fluent in both Lebanese dialect and English, evaluated each translation. The annotation process and the scoring rubric are provided in Appendix C.2.

Correlation Analysis: We calculated Krippendorff's alpha to measure the agreement between annotators. The threshold for acceptable agreement was set at $\alpha \geq 0.6$, indicating substantial agreement.

For each metric, we calculated:

- Pearson correlation coefficient (r) for linear correlation
- Spearman correlation coefficient (ρ) for monotonic correlation
- Statistical significance (p-value < 0.05)

The results of our assessment, presented in Table 2, reveal significant variations in metric performance. BLEU demonstrates the weakest alignment with human judgment, exhibiting minimal correlation coefficients (r = 0.098, ρ = 0.074). In contrast, xCOMET achieves a substantially higher correlation with human evaluations (r = 0.606, ρ = 0.631), indicating its superior reliability as an automatic evaluation metric. These findings underscore the comparative advantage of neural-based metrics, particularly COMET and xCOMET, over traditional approaches. Notably, the stronger performance of xCOMET compared to COMET may be attributed to its enhanced interpretability and larger model capacity. Furthermore, the results empirically demonstrate the limitations of BLEU as a reliable metric for translation quality assessment in this context.

Metric	r	ρ	p
BLEU	0.098	0.074	0.0336
BertScore	0.492	0.430	0.0000
COMET	0.523	0.461	0.0000
xCOMET	**0.606**	**0.631**	0.0000

Table 2: Correlation coefficients (Pearson's r and Spearman's ρ), measuring alignment between human scores and automated metrics

[6] https://chatgpt.com/

[7] https://dashboard.cohere.com/playground/chat

9 Conclusion

Unlike existing datasets derived from translated foreign sources, we curated, in this work, the Language Wave (LW) dataset that captures the nuances of colloquial Lebanese Arabic. Our linguistic analysis demonstrates LW's superior cultural richness, providing a resource that potentially aids the development of culturally sensitive AI applications.

Furthermore, our analysis reveals a striking disparity in model performance between non-native/translated and culturally-rich content, highlighting the inadequacy of current evaluation approaches for handling culturally nuanced content. In addition, we show the substantial performance gap between LLMs and encoder-decoder models when translating culturally relevant Lebanese content. While traditional encoder-decoder models often default to literal translations that fail to capture cultural significance, LLMs are usually better at finding cultural alternatives.

A comprehensive qualitative analysis of idiomatic expressions, cultural semantics embedded in Lebanese Arabic, and the inherent linguistic ambiguity of Arabic scripts highlights the complexity of translating Lebanese, a language deeply rooted in its culture. Finally, we demonstrate how this analysis can be adapted to other Arabic-script languages that share similar linguistic and cultural characteristics.

10 Limitations and Future Works

The current study presents some limitations. We evaluated LLMs only in a zero-shot setting, while there is a promising potential for exploring more sophisticated prompting techniques to enhance LLMs translation performance. The use of xCOMET score as an evaluation metric also can present limitations due to its Western-centric training data, indicating the need for more culturally appropriate evaluation methodologies, potentially through human evaluation or LLM-based assessment. While conducting the human assessment, we did not explicitly give instructions to score the fidelity of preserving cultural terms, and idioms in translation. While qualitative analysis provided valuable insights, a more comprehensive human evaluation remains an area for further exploration. Furthermore, while the Language Wave dataset represents a significant step forward, it does not fully capture the regional dialectal variations within Lebanon, and significant challenges remain in developing robust culturally-aware translation data, and accurately benchmarking these datasets. Finally, resource constraints limited our model evaluation scope, leaving several prominent multilingual LLMs untested, including Claude, LLaMA, and ALLaM (Bari et al., 2024).

The results of this work suggest that the path forward for the translation of Arabic-scripts low-resource languages may lie not just in scaling existing architectures, but in fundamentally rethinking how we approach cultural preservation, through the careful curation of culturally authentic training data and the potential advantages of open-source LLMs for handling culturally nuanced content. By demonstrating in this paper some of the limitations that LLMs face in translating Ajami scripts, we pave the way for the research community to explore the interplay between linguistic diversity and cultural preservation in translation.

11 Ackowledgments

We thank the Maroun Semaan Faculty of Engineering and Architecture (MSFEA) and the AI, Data Science, and Computing Hub (AIDSC) at the American University of Beirut for supporting this work through the Artificial Intelligence Exploratory Research Grant.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur A. Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Yousseif Elshahawy, Ahmed M. Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2023. Larabench: Benchmarking arabic ai with large language models. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Wael Abid. 2020. The sadid evaluation datasets for low-resource spoken language machine translation of arabic dialects. In *International Conference on Computational Linguistics*.

Rania Al-Sabbagh. 2023. The negative transfer effect on the neural machine translation of egyptian arabic adjuncts into english: The case of google translate. *International Journal of Arabic-English Studies.*

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. LLMs for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, St. Julian's, Malta. Association for Computational Linguistics.

Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. 2024. Dallah: A dialect-aware multimodal large language model for arabic. *ArXiv*, abs/2407.18129.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2024. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

Saikat Barua. 2024. Exploring autonomous agents through the lens of large language models: A review. *ArXiv*, abs/2404.04442.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The madar arabic dialect corpus and lexicon. In *International Conference on Language Resources and Evaluation.*

Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Improving llm-based machine translation with systematic self-correction. *ArXiv*, abs/2402.16379.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *ArXiv*, abs/2302.07856.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luísa Coheur, Pierre Colombo, and André Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Karim El Haff, Mustafa Jarrar, Tymaa Hammouda, and Fadi A. Zaraket. 2022. Curras + baladi: Towards a levantine corpus. In *International Conference on Language Resources and Evaluation.*

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

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

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Joy Ishaku, Muhammad Mustapha, and Muhammad Bello. 2020. Contrastive analysis of lexical and structural ambiguity between hausa and english languages. 2:19–34.

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

Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties. *Preprint*, arXiv:2308.03051.

Md. Tawkat Islam Khondaker, Numaan Naeem, Fatimah Khan, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Benchmarking llama-3 on arabic language generation tasks. In *ARABICNLP.*

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Annual Meeting of the Association for Computational Linguistics.*

Mateusz Krubiński, Hashem Sellat, Shadi Saleh, Adam Pospíil, Petr Zemánek, and Pavel Pecina. 2023. Multi-parallel corpus of north levantine arabic. In *ARABICNLP.*

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *WMT@ACL.*

Seungjun Lee, Jungseob Lee, Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, Seonmin Koo, and Heu-Jeoung Lim. 2023. A survey on evaluation metrics for machine translation. *Mathematics*.

Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, Benyou Wang, and Jinchao Xu. 2024. Alignment at pretraining! towards native alignment for arabic LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chenyang Lyu, Jitao Xu, Longyue Wang, and Minghao Wu. 2023. A paradigm shift: The future of machine translation lies with large language models. In *International Conference on Language Resources and Evaluation*.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur A. Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *ArXiv*, abs/2409.11404.

Tarek Naous, Michael Joseph Ryan, and Wei Xu. 2023. Having beer after prayer? measuring cultural bias in large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Fallou Ngom, Daivi Rodima-Taylor, and David Robinson. 2023. ʿajamī literacies of africa: The hausa, fula, mandinka, and wolof traditions. *Islamic Africa*, 14(2):119 – 143.

Viktória Ondrejová and Marek Šuppa. 2024. Can LLMs handle low-resource dialects? a case study on translation and common sense reasoning in šariš. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 130–139, Mexico City, Mexico. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *WMT@EMNLP*.

Ricardo Rei, José G. C. de Souza, Duarte M. Alves, Chrysoula Zerva, Ana C. Farinha, T. Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Conference on Machine Translation*.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, Alham Fikri Aji, Zhiqiang Shen, Zhengzhong Liu, Natalia Vassilieva, Joel Hestness, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Hector Xuguang Ren, Preslav Nakov, Timothy Baldwin, and Eric Xing. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *Preprint*, arXiv:2308.16149.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. *ArXiv*, abs/2309.16575.

Nllb team, Marta Ruiz Costa-jussà, James Cross, Onur cCelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Alison Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon L. Spruit, C. Tran, Pierre Yves Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzm'an, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *ArXiv*, abs/2207.04672.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *ArXiv*, abs/2309.11674.

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

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyridon Matsoukas, Richard M. Schwartz, John Makhoul, Omar Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *North American Chapter of the Association for Computational Linguistics*.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024. Teaching large language models an unseen language on the fly. *ArXiv*, abs/2402.19167.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

A Translation Models

Jais: MBZUAI introduced the largest openly available Arabic language models, known as Jais ranging from 590M to 70B (Sengupta et al., 2023), which quickly captured the attention of the Arabic research community. These models were built based on the GPT architecture and pre-trained using a blend of English and Arabic datasets, making them ideal candidates for the translation task. However, Jais's primary limitation lies in its heavy reliance on translated datasets, driven by the scarcity of high-quality Arabic datasets. Notably, this reliance on translated data can introduce ”localization issues,” potentially undermining the reliability and applicability of the models in native contexts, specifically in the translation of cultural content. (Huang et al., 2024) have observed an apparent bias in Jais, and showed that Jais produced outputs with a notable inclination toward English-centric content, frequently emphasizing terms associated with Christianity, for instance.

AceGPT: Decoder-only models built on top of LLaMA2, ranging from 7 billion to 70B billion(Liang et al., 2024)(Huang et al., 2024). Developers of AceGPT tried to address the challenge of Arabic localization and cultivate culturally and value-aligned Arabic LLMs capable of accommodating the diverse, application-specific needs of Arabic-speaking communities. They delved into the critical necessity and the methodology behind creating a localized Large Language Model specifically tailored for the Arabic language, which possesses distinct cultural traits that aren't adequately accommodated by current open-source mainstream models. Their main contribution was in using Reinforcement Learning with AI Feedback (RLHF) to align the model's responses with the cultural and value norms of Arabic-speaking communities. GPT4 was used to rank answers based on how well they represent Arabic values. Nonetheless, the Arabic localization challenge persists. The pool of prompts that Arabic users will use is pretty much different than the one used by English speakers, and it should predominantly reflect the queries of Arab users, which would inherently carry more cultural relevance.

Cohere Models: The Aya initiative, developed by CohereAI, seeks to bridge the gap between multilingual and monolingual model performance. Most promising multilingual Aya models are Aya23 and Aya-expanse which ranges from 8B to 33B parameters and cover 23 languages. Since its inception two years ago, the Aya project has involved a participatory research effort with over 3,000 contributors from 119 countries, fostering the development of culturally-aware AI. This collaboration has produced the largest multilingual dataset collection to date, consisting of 513 million examples, alongside comprehensive evaluation sets focused on multilingual performance and safety. In addition, the largest model from Cohere is Command-R+, a 104B parameter model with highly advanced capabilities, evaluated on 10 languages including Arabic. Unlike approaches that rely on translating English instruction-style datasets—prone to translation biases and loss of cultural context, Cohere's methodology emphasizes human-curated data collected through the Aya Annotation Platform. This platform facilitated the creation of the Aya dataset, which stands as the largest human-curated multilingual instruction finetuned dataset, enhancing the model's ability to reflect diverse cultural nuances and reducing the noise and biases typically associated with automatic dataset curation. As such, Cohere models are one of the most promising models to test on cultural understanding, especially in the translation of low-resource dialects.

NLLB Models: The NLLB (No Language Left Behind) project(team et al., 2022), launched by Meta AI in 2022, represents a significant leap forward in multilingual machine translation. This family of models ranging from 560M to 54B, is designed to support translations across 202 different language varieties, addressing the need for more inclusive language representation and overcoming the limitations that many models face when working with low-resource languages. Central to the NLLB project is its encoder-decoder architecture, which distinguishes it from large language models (LLMs) that primarily rely on decoder-only or transformer-based approaches. Unlike LLMs which are typically optimized for a broad range of generative tasks, the NLLB model's architecture is specifically tailored to translation, enabling more precise handling of input and output sequences. To ensure the quality of its translations, Meta AI introduced a comprehensive evaluation dataset called FLORES-200, which serves as a benchmark for assessing performance across all supported languages, and it showed NLLB superiority compared to existing datasets.

B Related Work

B.1 Benchmarking LLMs for Translation of Low-Resource/Dialectal Languages

The recent surge of Multilingual Large Language Models (MLLMs) has sparked a debate on their effectiveness in machine translation tasks compared to specialized translation systems(Xu et al., 2023). Research in (Hendy et al., 2023) and (Jiao et al., 2023) show that GPT models can translate effectively with proper prompting, however, they may struggle with specialized content in certain language pairs compared to dedicated translation services. Furthermore, studies have shown enhanced translation performance of open-source LLMs through better prompting, like self-correction (Feng et al., 2024), Dictionary-based prompting(Ghazvininejad et al., 2023), and imitating human-like thinking by splitting the translation task into small subtasks(He et al., 2023). Autonomous Agents were also explored in LLMs(Barua, 2024)

Despite these advancements, the issue of translating low-resource languages remains largely unaddressed. Both (Tanzer et al., 2023) and (Zhang et al., 2024) show that LLMs are capable of translating a new language that did not exist in the pre-training data. A paper that discusses how they leveraged LLMs for translation of low-resource languages in Saris (Ondrejová and Šuppa, 2024).

B.2 Benchmarking LLMs on Arabic translation

In the domain of machine translation (MT) from Arabic dialects to English, significant advancements have been made through the development of specialized datasets and the use of pre-trained Neural networks. Despite these efforts, the scarcity of parallel corpora for less common Arabic dialects and English poses a challenge, with most neural machine translation systems, including Google Translate, primarily relying on MSA and English corpora. This approach has proved its weakness, as evidenced by the authors in (Al-Sabbagh, 2023) who evaluated Google Translate's performance in the Egyptian dialect. Researchers in https://aclanthology.org/2024.arabicnlp-1.24.pdf benchmarked LLaMA3 on NLG Arabic tasks, including translation of code-switched arabic dialects to English. (Kadaoui et al., 2023) focused on evaluating the capabilities of models such as Bard and ChatGPT across a spectrum of Arabic dialects. They evaluated NLLB as the supervised baseline, finding both ChatGPT and GPT-4 able to outperform this baseline in a zero-shot setting. Still, this research underscores the challenges related to dialectal diversity and linguistic inclusivity of the Lebanese dialect and only evaluates large closed models. Superior LLMs like ChatGPT and GPT-4 are only accessible through restricted APIs, which creates barriers to new research and advancements in the field. None of these works focused on evaluating dialectal MT tasks for Smaller Arabic language models such as AceGPT and Jais. (Khondaker et al., 2024) benchmarked LLaMA3 on NLG Arabic tasks, including translation of code switched Arabic dialects to English. (Abdelali et al., 2023) developed LAraBench, a benchmarking Arabic AI with Large Language Models, they benchmarked on the AraBench. Likewise, (Abid, 2020) developed the SADID benchmark for evaluating Arabic dialects. However, they asked people what are the most topics they speak in their dialect, and they selected sources from Wikipedia in English, and then translated them. however, they chose English as the language of our source sentences instead of MSA so as not to bias our translations.

B.3 LLMs and cultural-awarness

Translating culture-related content is vital for effective cross-cultural communication. Recent research has benchmarked machine translation for cultural awareness (Yao et al., 2024) and demonstrated that Large Language Models (LLMs) exhibit superior capabilities compared to traditional neural MT systems in leveraging external cultural knowledge, especially for Culturally-Specific Items (CSIs) translation. In the Arabic language domain, this challenge is further complicated by dialectal variations and the scarcity of high-quality datasets. This difficulty hinders the analysis of cultural awareness of machine translation (MT) systems, including traditional neural MT and the emerging MT paradigm using large language models (LLM). Arabic-centric LLMs like Jais and AceGPT, while showing promise in Arabic NLP, face limitations due to their reliance on translated datasets, introducing ”localization issues”(Huang et al., 2024). Recent initiatives like Dallah(Alwajih et al., 2024), a dialect-aware multimodal LLM for Arabic, represent ongoing efforts to better accommodate the distinct cultural traits and dialectal variations that current mainstream models struggle to capture. Nevertheless, some effort have been made to benchmark LLMs on cultural awareness. (Naous et al., 2023) measured the cultural bias and LLMs , while AraDICE benchmark(Mousi et al., 2024) was developed to assess LLMs' cultural awareness and dialect comprehension. Researchers leveraged MT, specifically from English to MSA and MSA to dialects, combined with human post-editing, to develop synthetic benchmarks for low-resource DA. However, these evaluation efforts themselves often rely on translated benchmarks from English to Modern Standard Arabic (MSA) and subsequently to dialects, highlighting a persistent challenge in developing authentic resources for low-resource Arabic dialects. While current work on cultural awareness in Arabic dialects primarily focuses on CSIs, the challenge extends far beyond isolated cultural items to encompass the entire linguistic system - including verbs, vocabulary, grammar structures, and idiomatic expressions that are deeply rooted in cultural context. Despite dialects being deeply rooted in cultural context, the field continues to rely heavily on translated data due to resource scarcity, suggesting a critical need to redirect efforts toward developing authentic, culturally-aware datasets that capture the full richness of Arabic dialectal variations.

C Aligning Metrics with Human Judgement

In the field of Neural Machine Translation (NMT), the accurate evaluation of translation quality remains a critical challenge. While traditional lexical-based metrics such as BLEU (Papineni et al., 2002) and CHRF(Popovic, 2015) have been widely used, they often fall short in capturing the nuanced aspects of translation quality, particularly semantic equivalence and grammatical correctness. This limitation has led to the development of more sophisticated evaluation techniques, among which xCOMET stands out as a promising solution.

C.1 Translation Evaluation Metrics

The evolution of machine translation metrics can be broadly categorized into four main types:

1. **Lexical-based metrics**: These include widely used measures such as BLEU(Papineni et al., 2002), METEOR(Lavie and Agarwal, 2007), and TER(Snover et al., 2006). While these metrics have been instrumental in the development of NMT systems, they primarily focus on surface-level similarities between the machine translation output and reference translations. Their inability to account for semantic equivalence limits their effectiveness in accurately assessing translation quality.

2. **Embedding-based metrics**: These metrics, such as BERTScore(Zhang et al., 2019), utilize contextual embeddings to capture semantic similarities between translations. By leveraging pre-trained language models, they offer a more nuanced evaluation that considers the context.

3. **Supervised metrics**: These metrics, exemplified by Cross-lingual Optimized Metric for Evaluation of Translation(COMET)(Rei et al., 2022), are trained on human judgments of translation quality. While they show a higher correlation with human evaluations, their reliance on labeled data can limit their applicability to low-resource languages.

4. **Interpretable metrics**: This emerging category of metrics aims to provide transparent and explainable evaluations of machine translations. xCOMET(Guerreiro et al., 2023) falls into this category, offering significant advantages over previous approaches. Unlike black-box metrics, xCOMET provides detailed insights into specific translation errors. This granular approach allows for a more comprehensive understanding of translation quality and pinpoints areas for improvement. It can also be used for quality estimation without a reference, reference-only evaluation, or full source-reference-hypothesis evaluation. This flexibility makes it a versatile tool for various translation assessment needs. By leveraging advanced language models and fine-grained error detection, xCOMET achieves a higher correlation with human evaluations compared to traditional metrics. With models ranging from 3.5B parameters (xCOMET-XL) to 10.7B parameters (xCOMET-XXL), xCOMET can be scaled to meet various computational requirements and evaluation needs.

C.2 Metric Correlation Analysis

Annotation Process: Each annotator independently rated all 150 translations. Annotations were collected through a spreadsheet with source text, translation, and scoring columns. Annotators were instructed to:

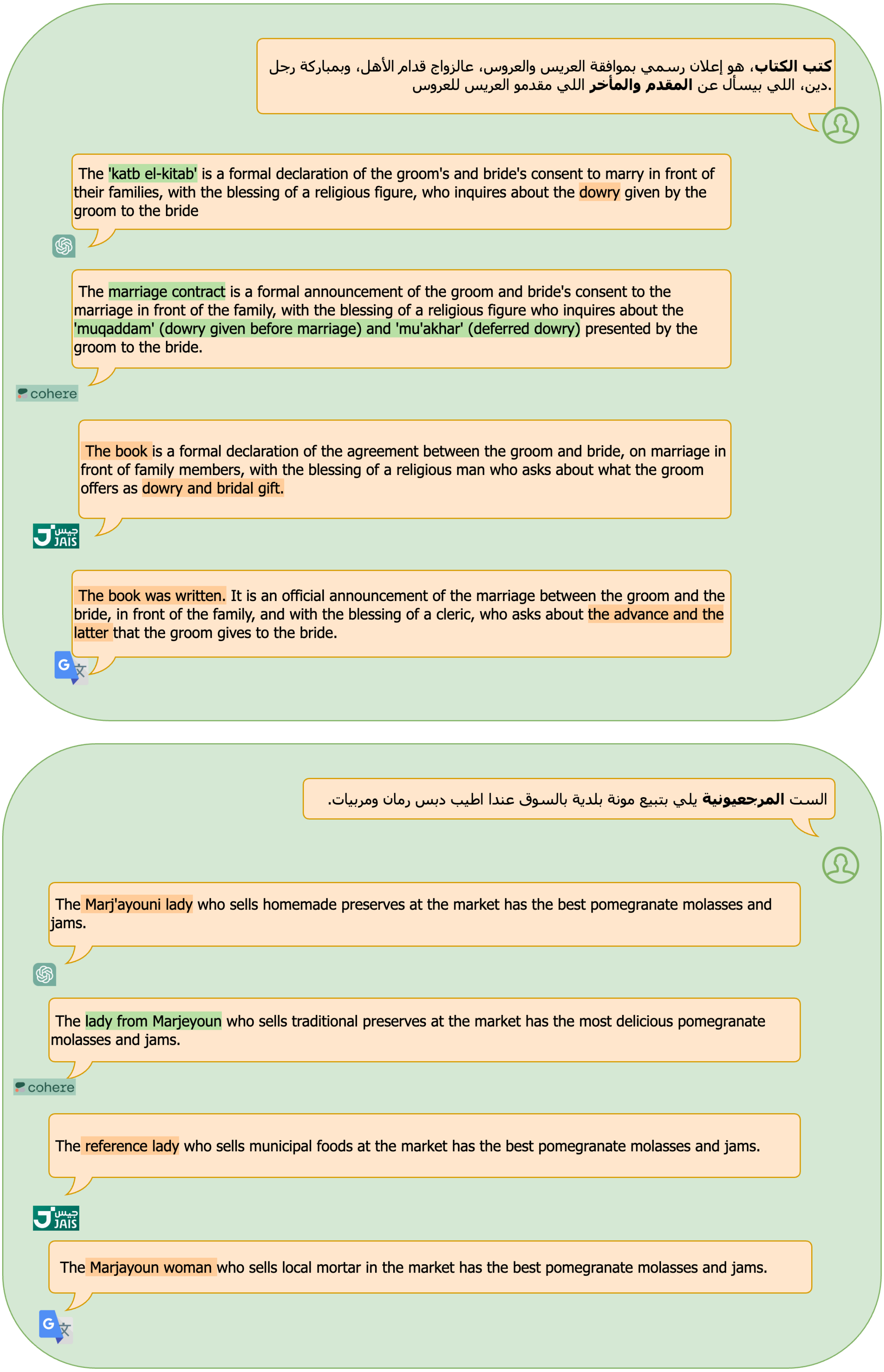
1. Read both source and translation carefully
2. Consider both accuracy and fluency
3. Apply scores consistently according to the rubric

Annotation Guidelines: We instructed annotators to carefully read and follow the guidelines shown in Table 3.

Score	Category	Description	Examples
5	Very Good	• Completely preserves meaning • Natural English expression • No grammatical errors	• Source: شو عم تعمل ؟ • Translation: What are you doing?
4	Good	• Minor flaws that don't affect understanding • Slight unnatural expressions • Minor grammatical issues	• Source: عم موت من البرد • Translation: I am dying from the cold • Comment: slightly literal but acceptable
3	Adequate	• Core meaning preserved • Some unnatural expressions • Notable but non-critical errors	• Source: شو هالحكي • Translation: What is this talk • Comment: understandable but unidiomatic
2	Poor	• Significant meaning loss • Major grammatical errors • Difficult to understand	• Source: عطيني نَفَس • Translation: Give me breath • Comment: literal translation
1	Incomprehensible	• Complete meaning loss • Severe grammatical errors • Impossible to understand	• Source: حلّو عن بعض ! • Translation: Sweet each other! • Comment: completely misses meaning

Table 3: Translation Quality Assessment Rubric for Lebanese Dialect to English Translation

D Qualitative Examples



كتب الكتاب، هو إعلان رسمي بموافقة العريس والعروس، عالزواج قدام الأهل، وبمباركة رجل
دين، اللي بيسأل عن **المقدم والمأخر** اللي مقدمو العريس للعروس.

The 'katb el-kitab' is a formal declaration of the groom's and bride's consent to marry in front of their families, with the blessing of a religious figure, who inquires about the dowry given by the groom to the bride

The marriage contract is a formal announcement of the groom and bride's consent to the marriage in front of the family, with the blessing of a religious figure who inquires about the 'muqaddam' (dowry given before marriage) and 'mu'akhar' (deferred dowry) presented by the groom to the bride.

The book is a formal declaration of the agreement between the groom and bride, on marriage in front of family members, with the blessing of a religious man who asks about what the groom offers as dowry and bridal gift.

The book was written. It is an official announcement of the marriage between the groom and the bride, in front of the family, and with the blessing of a cleric, who asks about the advance and the latter that the groom gives to the bride.

الست **المرجعيونية** يلي بتبيع مونة بلدية بالسوق عندا اطيب دبس رمان ومربيات.

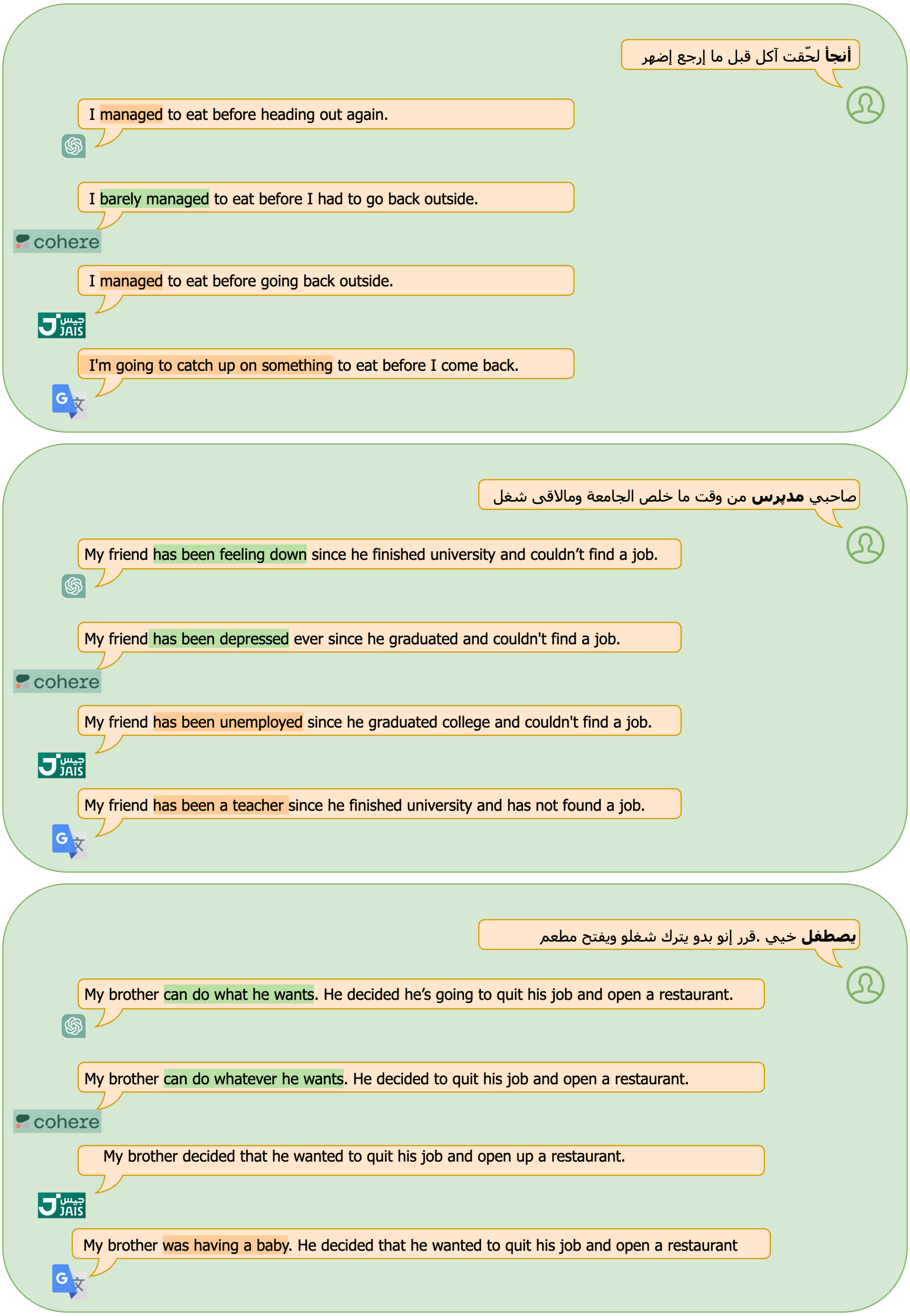
The Marj'ayouni lady who sells homemade preserves at the market has the best pomegranate molasses and jams.

The lady from Marjeyoun who sells traditional preserves at the market has the most delicious pomegranate molasses and jams.

The reference lady who sells municipal foods at the market has the best pomegranate molasses and jams.

The Marjayoun woman who sells local mortar in the market has the best pomegranate molasses and jams.

Figure 3: Two examples highlighting the performance of four models: Jais-70B, Command-R+, GPT-4o and GoogleTranslate in translating Lebanese cultural expressions. The first example contains social terms used in a Lebanese Wedding, while the second example refers to a Lebanese custom in one village. **Bold**: Challenging Lebanese Terms 🟩: correct translation 🟧: wrong translation



أنجأ لحّقت آكل قبل ما إرجع إضهر

I managed to eat before heading out again.

I barely managed to eat before I had to go back outside.

I managed to eat before going back outside.

I'm going to catch up on something to eat before I come back.

صاحبي **مدپرس** من وقت ما خلص الجامعة ومالاقى شغل

My friend has been feeling down since he finished university and couldn’t find a job.

My friend has been depressed ever since he graduated and couldn't find a job.

My friend has been unemployed since he graduated college and couldn't find a job.

My friend has been a teacher since he finished university and has not found a job.

يصطفل خيي .قرر إنو بدو يترك شغلو ويفتح مطعم

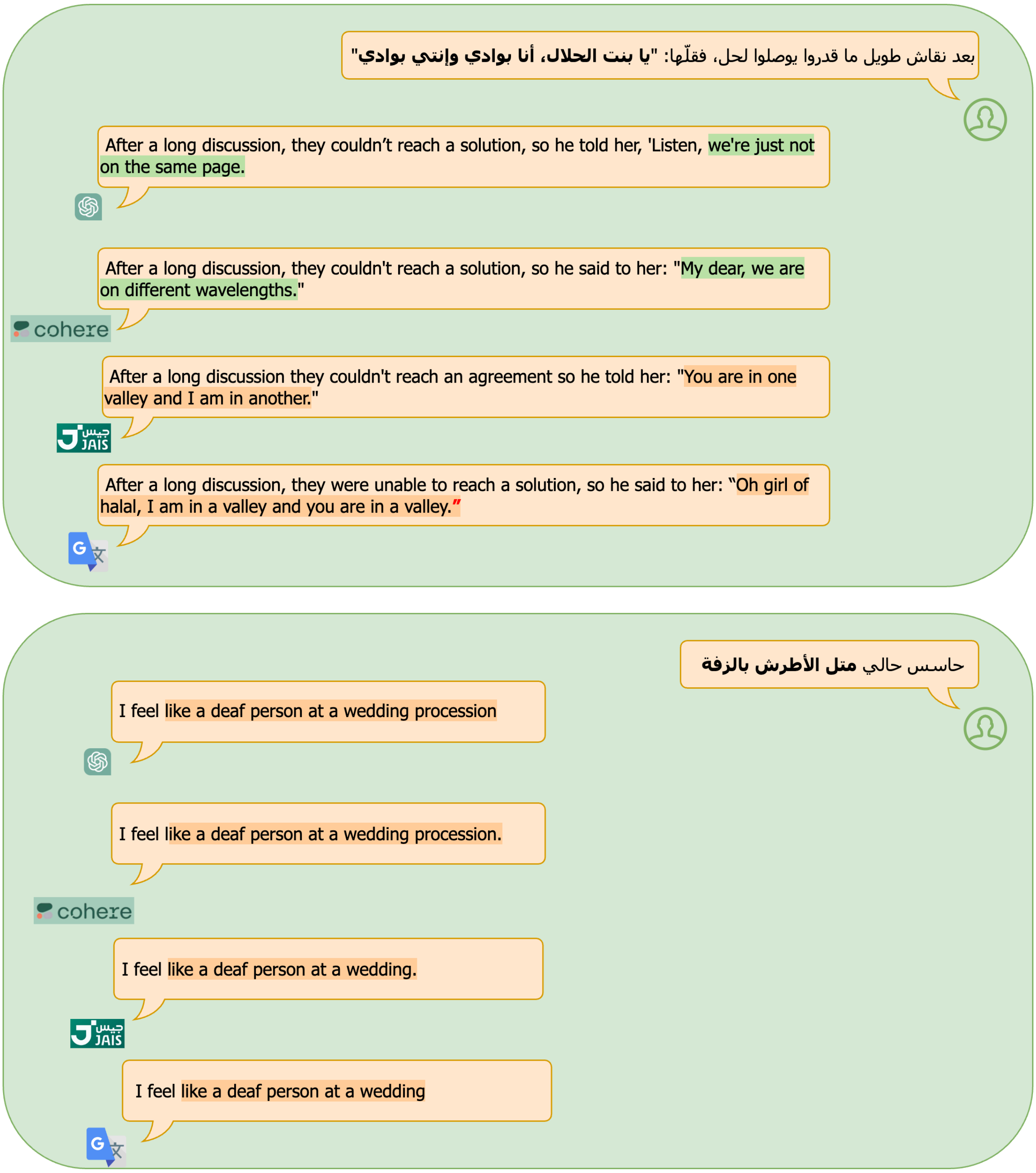
My brother can do what he wants. He decided he’s going to quit his job and open a restaurant.

My brother can do whatever he wants. He decided to quit his job and open a restaurant.

My brother decided that he wanted to quit his job and open up a restaurant.

My brother was having a baby. He decided that he wanted to quit his job and open a restaurant

Figure 4: Three examples showing the performance of four models: Jais-70B, Command-R+ from Cohere, GPT-4o and GoogleTranslate in translating unique Lebanese linguistic terms. The first example contains the Lebanese term أنجأ, the second example have the Lebanized word (مدپرس), while the third example focuses on the translation of the famous Lebanese word (يصطفل).
Bold: Challenging Lebanese Terms 🟩: correct translation 🟧: wrong translation



بعد نقاش طويل ما قدروا يوصلوا لحل، فقلّها: "**يا بنت الحلال، أنا بوادي وإنتي بوادي**"

After a long discussion, they couldn't reach a solution, so he told her, 'Listen, we're just not on the same page.

After a long discussion, they couldn't reach a solution, so he said to her: "My dear, we are on different wavelengths."

After a long discussion they couldn't reach an agreement so he told her: "You are in one valley and I am in another."

After a long discussion, they were unable to reach a solution, so he said to her: "Oh girl of halal, I am in a valley and you are in a valley."

حاسس حالي **متل الأطرش بالزفة**

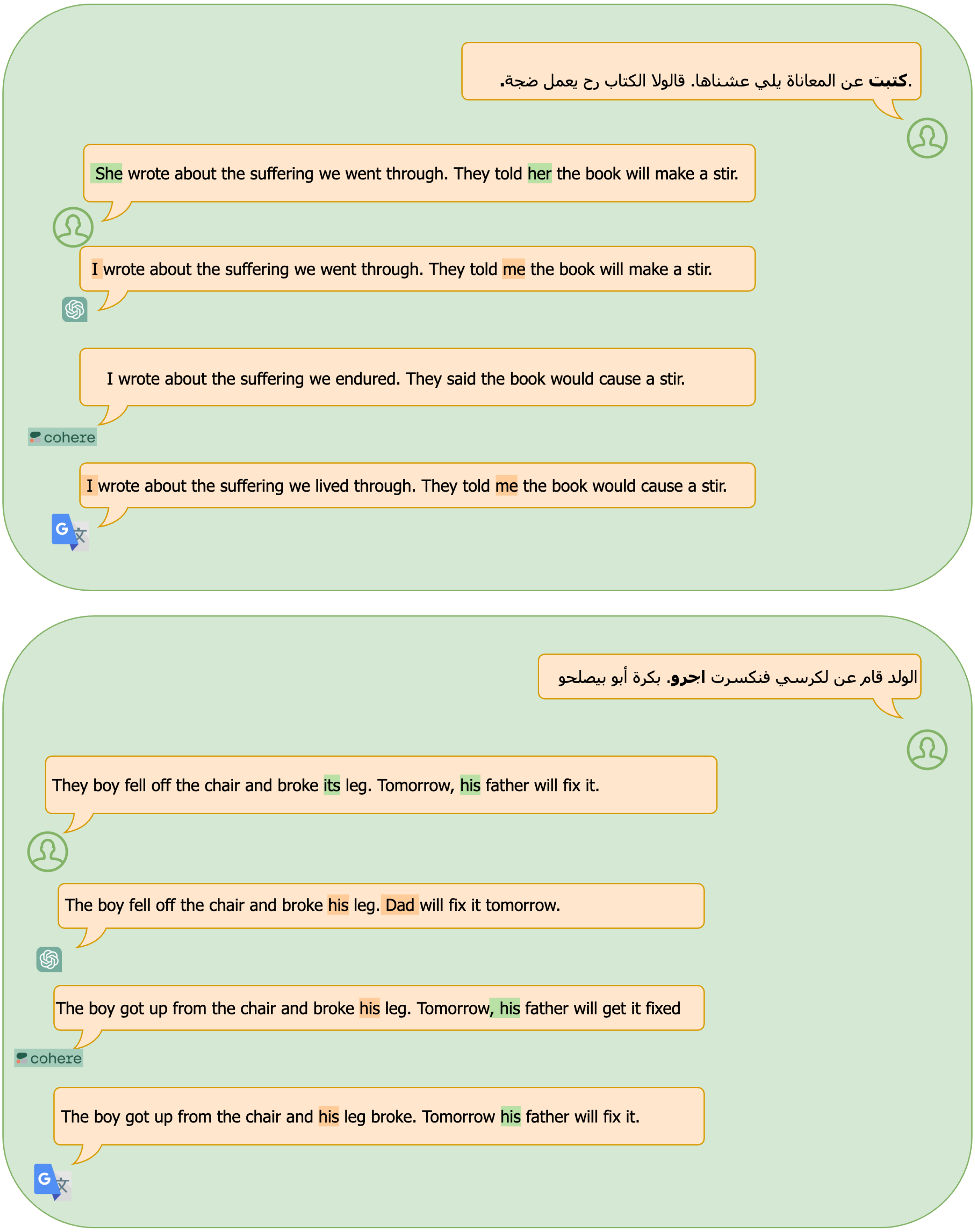
I feel like a deaf person at a wedding procession

I feel like a deaf person at a wedding procession.

I feel like a deaf person at a wedding.

I feel like a deaf person at a wedding

Figure 5: Two examples focusing on the performance of four models: Jais-70B, Command-R+, GPT-4o and GoogleTranslate in translating Lebanese idioms. The first example depicts a famous idiom وإنتي بوادي أنا بوادي which means "We're on different pages", while the second example shows the proverb (حاسس متل الأطرش بالزفة) which means "I feel out of place".
Bold: Challenging Lebanese Terms 🟩: correct translation 🟧: wrong translation



كتبت عن المعاناة يلي عشناها. قالولا الكتاب رح يعمل ضجة.

She wrote about the suffering we went through. They told her the book will make a stir.

I wrote about the suffering we went through. They told me the book will make a stir.

I wrote about the suffering we endured. They said the book would cause a stir.

I wrote about the suffering we lived through. They told me the book would cause a stir.

الولد قام عن لكرسي فنكسرت **اجرو**. بكرة أبو بيصلحو

They boy fell off the chair and broke its leg. Tomorrow, his father will fix it.

The boy fell off the chair and broke his leg. Dad will fix it tomorrow.

The boy got up from the chair and broke his leg. Tomorrow, his father will get it fixed

The boy got up from the chair and his leg broke. Tomorrow his father will fix it.

Figure 6: Two examples focusing on the performance of three models: Command-R+, GPT-4o and GoogleTranslate in translating Lebanese ambiguous expressions. The first example depicts the verbكتبت which can either mean ”I wrote” or ”she wrote”, while the second example show the expression اجرو which can translate into ”his leg” or ”its leg”.
Bold: Challenging Lebanese Terms 🟩: correct translation 🟧: wrong translation

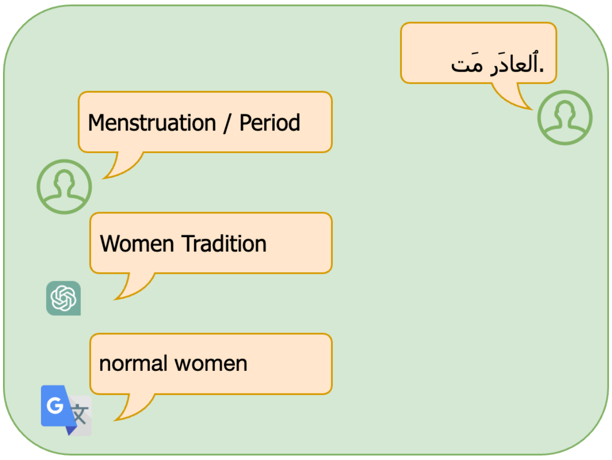


Figure 7: Example showing the translation of GPT-4o and Google-Translate for the Hausa expression **”al’adar mata”** (ألْعادَر مَت), a cultural term that refers to the women menstruation. The word **”mata”**(مَت) in Hausa means tradition but when talking about women, it refers to the monthly menstrual cycle, thus *GPT-4o* literally translated the expression to ”Women Traditions”.

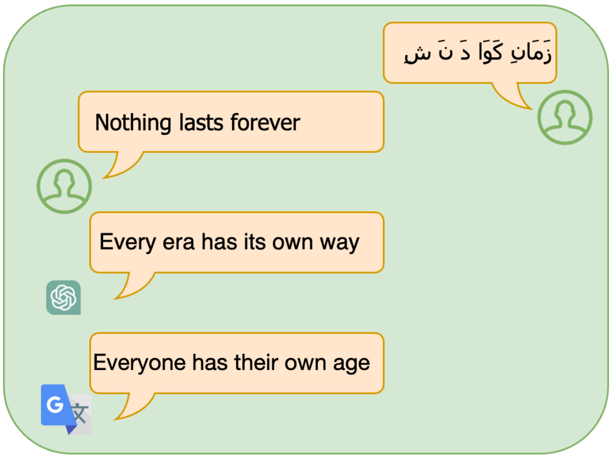


Figure 8: Example showing the translation of GPT-4o and Google-Translate for the Hausa proverb **”Zamani kowa da na shi”**(زَمَانِ كَوَا دَ نَ شِ) which literally translates to ”Everyone has his reign”. The proverb is used to mean that nothing lasts forever. It also refers to the fact that each regime comes with its policies, which will not last forever.

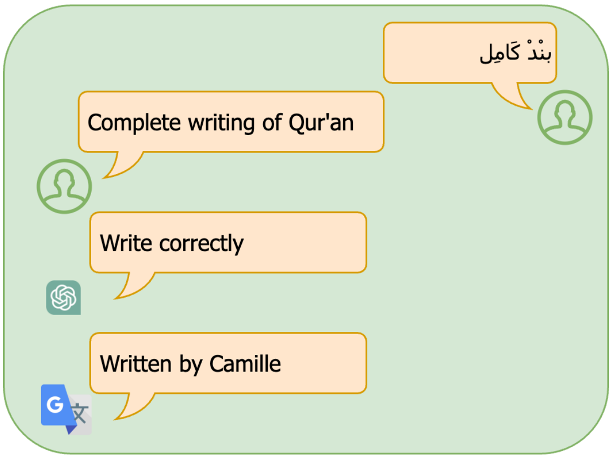


Figure 9: Example showing the translation of GPT-4o and Google-Transalte of the Wolof expression **"Bind Kamiil"**(بِنْدْ كَامِل), an expression term that refers to the practice of "writing an entire copy of the Quran", before graduating from the elementary level of Quranic education. GPT-4o and Google Translate fail to acknowledge the cultural relevance of this expression.

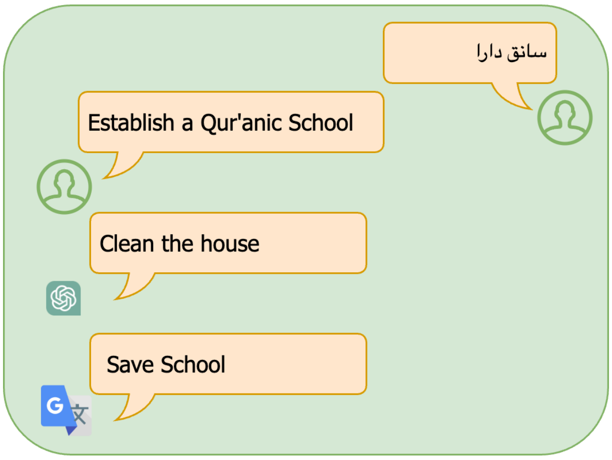


Figure 10: Example showing the translation of GPT-4o and Google-Translate for the Wolof term **"Sànc daara"**(سانق دارا), a religious expression that means "To create a Quranic school". It is regarded as an honor in Wolof society and one of the ultimate goals of many Quranic school students. While (سانق) can have many meanings clean/establish/save, using (سانق دارا) together usually refers to building a Qur'anic school.