Jawaher: A Multidialectal Dataset of Arabic Proverbs for LLM Benchmarking

Samar M. Magdy[ξ*] Sang Yun Kwon[λ*] Fakhraddin Alwajih[λ]
Safaa Abdelfadil[ξ] Shady Shehata[ξ,γ] Muhammad Abdul-Mageed[λ,ξ,γ]
[λ]The University of British Columbia [ξ]MBZUAI [γ]Invertible AI
samar.magdy@mbzuai.ac.ae, {skwon01, muhammad.mageed}@ubc.ca

Abstract

Recent advancements in instruction fine-tuning, alignment methods such as reinforcement learning from human feedback (RLHF), and optimization techniques like direct preference optimization (DPO), have significantly enhanced the adaptability of large language models (LLMs) to user preferences. However, despite these innovations, many LLMs continue to exhibit biases toward Western, Anglo-centric, or American cultures, with performance on English data consistently surpassing that of other languages. This reveals a persistent cultural gap in LLMs, which complicates their ability to accurately process culturally rich and diverse figurative language, such as proverbs. To address this, we introduce *Jawaher*, a benchmark designed to assess LLMs' capacity to comprehend and interpret Arabic proverbs. *Jawaher* includes proverbs from various Arabic dialects, along with idiomatic translations and explanations. Through extensive evaluations of both open- and closed-source models, we find that while LLMs can generate idiomatically accurate translations, they struggle with producing culturally nuanced and contextually relevant explanations. These findings highlight the need for ongoing model refinement and dataset expansion to bridge the cultural gap in figurative language processing. Project GitHub page is accessible at: https://github.com/UBC-NLP/jawaher.

1 Introduction

Instruction fine-tuning (Chung et al., 2024) has significantly enhanced the creativity and customizability of LLMs, while alignment techniques, such as RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2024), have improved the ability of these models to align with user preferences. The vast reservoir of cultural knowledge embedded within LLMs, combined with the potential of these alignment techniques, has sparked interest in how LLMs can reflect specific human values and personas across different cultures (Gupta et al., 2023; Kovač et al., 2023).

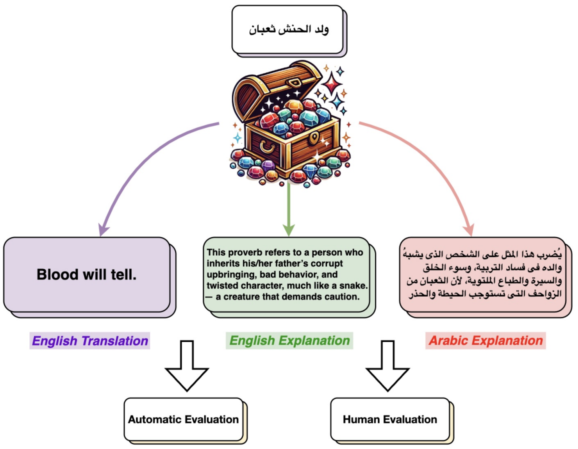

ولد الحنش ثعبان
Blood will tell.
This proverb refers to a person who inherits his/her father's corrupt upbringing, bad behavior, and twisted character, much like a snake. — a creature that demands caution.
يُضرب هذا المثل على الشخص الذى يشبهُ والده فى فساد التربية، وسوء الخلق والسيرة والطباع الملتوية، لأن الثعبان من الزواحف التى تستوجب الحيطة والحذر
English Translation
English Explanation
Arabic Explanation
Automatic Evaluation
Human Evaluation

Figure 1: Overview of the *Jawaher* benchmark, featuring Arabic proverbs along with their idiomatic translations and explanations in both English and Arabic; for instance, the *Libyan* proverb ولد الحنش ثعبان whose literal meaning is "the son of a snake is a serpent", is explained as conveying the idea that "like father, like son in terms of bad behavior and twisted character".

Figurative (i.e., non-literal) language (Fussell and Moss, 2008; Shutova, 2011; Gibbs and Colston, 2012) rich with implicit cultural references, varies significantly across cultures and relies on shared cultural knowledge. Differences in figurative expressions between languages often arise from unique cultural values, historical context, and other regional factors. Thus, understanding figurative language depends on grasping culturally significant concepts and their associated sentiments (Kabra et al., 2023). To ensure LLMs are inclusive and deployable across regions and applications, it is essential for these models to function adequately in diverse cultural contexts (Adilazuarda et al., 2024). Among common types of figurative language, *proverbs* are a key form of

[*] Equal contribution

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 12320–12341
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

cultural expression, encapsulating diverse areas of human knowledge and experience (González Rey, 2002; Mieder, 2004). Despite their widespread use, proverbs are linguistically complex due to their cultural significance and structure, exhibiting notable lexical and syntactic variations (Chacoto, 1994), making their automatic identification in text challenging for NLP systems.

Recent research aiming to study cultural representation and inclusion in LLMs has found that many models remain strongly biased towards Western, Anglo-centric, or American cultures, with performance on English data still surpassing that of other languages (Dwivedi et al., 2023; Adilazuarda et al., 2024). This highlights the existence of significant '*culture gaps*' (Liu et al., 2023) in LLMs, which further complicate their ability to accurately handle culturally rich and diverse forms like proverbs. In this paper, we make several contributions toward closing the cultural gap in LLMs by introducing *Jawaher*, which facilitates the building and assessment of models designed to understand figurative language, specifically Arabic proverbs, as follows:

(i) Jawaher: a new benchmark for figurative language understanding of Arabic proverbs. *Jawaher* comprises 10,037 high-quality Arabic proverbs with multidialectal coverage across 20 varieties, offering a rich and diverse collection representative of a wide range of Arab countries (Section 3). The proverbs are paired with idiomatic translations and explanations in English (Section 3.2) that cover different themes and cultural context (Section 3.1). Figure 1 illustrates the pairing of proverbs in *Jawaher* with their translation and explanations.

(ii) Comprehensive experiments to assess *Jawaher*'s usefulness towards building models that understand Arabic proverbs. We conduct a comprehensive set of experiments using both *open*- and *closed*- source models to test their abilities in interpreting, contextualizing, and translating these proverbs (Section 4.1). We propose extensive automatic and human evaluation to assess model understanding across our proposed tasks (Section 4.2). We find that while models generally perform well on *translation* tasks—producing idiomatically correct outputs—they struggle significantly with *explanation* tasks, particularly in capturing the cultural nuances, historical context, and deeper figurative meanings behind Arabic proverbs. Although closed-source models outperform open-source models, both still face notable limitations in clarity, cultural relevance, and detail in explanations, revealing a substantial gap in fully understanding and conveying the richness of Arabic proverbs (Section 5).

2 Related Works

Figurative Language. Prior work on figurative language understanding has covered a wide range of topics, including simile detection and generation (Niculae and Danescu-Niculescu-Mizil, 2014; Mpouli, 2017; Zeng et al., 2020). Results show that, although language models can often recognize non-literal language and may attend to it less, they still struggle to fully capture the implied meaning of such phrases (Shwartz and Dagan, 2019). This has led to more recent studies shifting toward tasks focused on *comprehending* figurative language (Chakrabarty et al., 2022; He et al., 2022; Prystawski et al., 2022; Jang et al., 2023). Efforts in dataset and task development have focused on diverse tasks such as recognizing textual entailment, multilingual understanding, and figurative language beyond text (Liu et al., 2023; Yosef et al., 2023; Saakyan et al., 2024). However, the lack of comprehensive datasets focusing on figurative language, particularly Arabic proverbs, limits the ability to thoroughly evaluate LLMs' cultural awareness and their true understanding of culturally embedded non-literal expressions, further motivating our work. More details regarding dataset and task development on figurative language are provided in Appendix A.

Arabic Proverbs. Proverbs are essential repositories of cultural values and historical experiences, conveying general truths or advice using non-literal language (Kuhareva, 2008; Brosh, 2013; Mieder, 2021). Despite their apparent simplicity, proverbs are complex linguistic and cultural products (Mieder, 2007), marked by distinct features that set them apart from ordinary language (Mieder, 2004; Tursunov, 2022). Arabic uniquely maintains a well-defined link between past and present in its linguistic and cultural traditions (Versteegh, 2014). Arabic proverbs, rooted in diverse dialects, reflect this continuity, showcasing not only the linguistic richness of the language but also the cultural, historical, and social values of different communities (Karoui et al., 2015; Elmitwally and Alsayat, 2020). More details on the cultural significance and linguistic complexity of Arabic proverbs are available in the Appendix B.

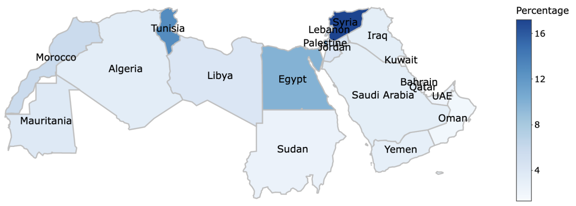


Figure 2: Choropleth map showing the geographical distribution of Arabic varieties covered in *Jawaher*. Color intensity represents the percentage of proverbs collected from each region, with darker shades indicating higher concentrations.

Cultural LLM. LLMs have attracted substantial interest in sociocultural studies, particularly regarding their performance across diverse cultural contexts (Gupta et al., 2023; Kovač et al., 2023). Research has increasingly uncovered a cultural gap, demonstrating that many models are biased toward Western, Anglo-centric perspectives (Johnson et al., 2022; Liu et al., 2023). These biases impact linguistic-cultural interactions and challenge value-based objectives (Johnson et al., 2022; Durmus et al., 2023). Efforts to address this include multilingual QA (Kabra et al., 2023), cross-cultural translation (Singh et al., 2024), and culturally diverse dataset creation (Ji et al., 2024; Qian et al., 2024), all aiming to improve multilingual adaptation and cultural alignment in LLMs.

3 Jawaher

Jawaher consists of Arabic proverbs paired with their idiomatic or literal English translations, along with explanations in both Arabic and English, covering 20 different Arabic varieties. Below, we outline the analysis of *Jawaher*, focusing on its coverage of dialects, themes, cultural context, and the tasks it facilitates.

3.1 *Jawaher* Analysis

Dialect Representation. Data in *Jawaher* is manually curated by four native Arabic speakers with strong linguistic expertise. The four annotators come from Egypt (two annotators), Mauritania, and Morocco. During data collection, they consulted with native speakers from other countries such as Jordan, Syria, and the United Arab Emirates (UAE) to ensure a diverse and authentic representation of proverbs from countries across the Arab world. We acquire data from publicly available online resources and carefully verify the origins of proverbs to confirm that they truly reflect their respective cultural heritage. Figure 2 shows the geographical distribution of the countries covered in *Jawaher*, highlighting the dataset's regional diversity. Modern Standard Arabic (MSA), a pan-Arabic variety not tied to any specific region, is excluded to better showcase the thematic distribution of proverbs across distinct Arab dialects and regions. Below is a list of all covered countries, categorized by dialect group and region: ***Gulf Dialects:*** Bahraini, Emirati, Kuwaiti, Omani, Qatari, and Iraqi; ***Levantine Dialects:*** Jordanian, Lebanese, Palestinian, and Syrian; ***North African Dialects:*** Algerian, Libyan, Mauritanian, Moroccan, and Tunisian; ***Arabian Peninsula:*** Yemeni and Saudi; and ***Nile Basin:*** Egyptian and Sudanese.

Theme Classification. Popular proverbs in Arab countries address various aspects of daily life, serving as an important source of folk culture while reflecting the historical, religious, and societal values of their users. These proverbs often reveal insights into the evolution of civilization and sometimes reference specific events or beliefs related to human existence. Dominant themes frequently incorporate elements such as animals, food, and other culturally significant symbols, to reflect the unique nature and heritage of Arab societies (Farghal, 2021).

We analyze the themes present in *Jawaher* by collecting embeddings of each proverb using a multilingual text embedding model[1] (Wang et al., 2024). and perform UMAP visualization on the collected data. Figure 3 reveals groupings representing different themes, such as food, mother, children, and eye. Upon inspection, these groupings not only bring sentences together by thematic content but also highlight cultural traits specific to each language or variety. *'Body parts'* emerges as a recurring theme, as seen in proverbs from Palestinian, Syrian, and Kuwaiti dialects, where references to '*eyes*' convey both literal meanings and deeper cultural values. The Omani proverb الأعور على العميان باشا (*Among the blind, the one-eyed is a king*) emphasizes how a person who has one flaw thinks that he is better than someone who has many flaws, as he is better than them. In the Syrian proverb عين ما بتقاوم مخرز (*An eye cannot resist an awl*), we see a metaphor for vulnerability,

[1] https://huggingface.co/intfloat/multilingual-e5-large

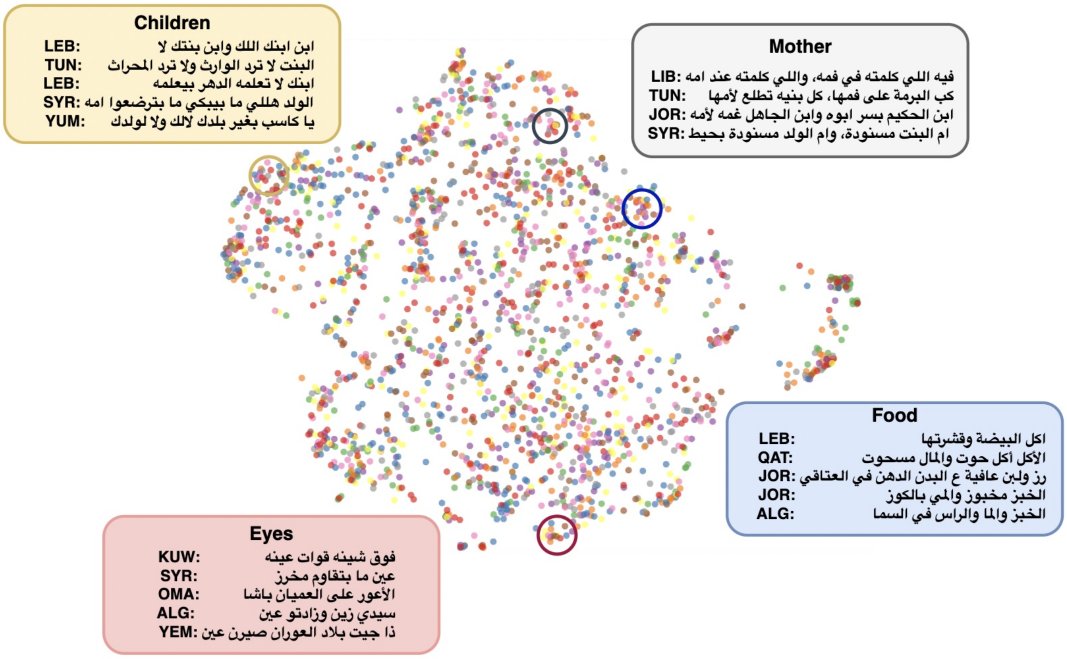


Figure 3: UMAP visualization of proverb embeddings, showing clusters corresponding to four key themes: *Food*, *Mother*, *Children*, and *Eyes*. Proverbs are grouped based on thematic and cultural similarities across various Arab dialects.

underscoring the inevitability of hardship. Meanwhile, the Kuwaiti proverb فوق شينه قوات عينه (*On top of his ugliness, he is brazen*) reflects disdain for arrogance despite flaws, emphasizing cultural disapproval of audacity without merit. These examples illustrate how proverbs portray unique cultural traits specific to each society through distinct linguistic expressions.

Cultural Context. Culture is an integral aspect of language and is crucial for understanding people's backgrounds and social interactions. Without this context, the nuanced messages and symbolic references in proverbs may be lost or misinterpreted (Webster, 2000). In structure mapping theory (Gentner, 1983), figurative language involves comparing a source and target concept, and Arab proverbs often use this mechanism to draw on shared cultural experiences across Arab societies (Kabra et al., 2023; Hamdan et al., 2023). The cultural context in Arabic proverbs is thus essential for interpreting their true meanings and appreciating their significance (Bakalla, 2023).

As illustrated above, certain themes reflect shared cultural experiences across the Arab world, while the expressions used to convey these themes vary across dialects. To further demonstrate the distinctiveness and connections between proverbs, we plot sentence-level representations using Kernel Density Estimate (KDE) (dimensionality reduction to two components using tSNE (van der Maaten and Hinton, 2008) using sentence embeddings described above. Figure 4 illustrates both shared cultural themes and regional variation across Arabic dialects. Notably, while the central dense areas in all subfigures reflect common cultural values, certain groups show distinct distributions. Proverbs from Gulf dialects, including *Bahraini* (BAH), *Kuwaiti* (KUW), and *Qatari* (QAT), exhibit substantial overlap, reflecting shared cultural and linguistic traditions within the region. Levantine dialects demonstrate strong cohesion, with *Jordanian* (JOR), *Lebanese* (LEB), and *Palestinian* (PAL) proverbs clustering closely, indicating deep cultural interconnections. North African dialects display greater dispersion, particularly among *Algerian* (ALG), *Mauritanian* (MAU), and *Tunisian* (TUN) proverbs, suggesting linguistic diversity within this region. The Arabic Peninsula and Nile Basin groups, represented by *Saudi* (SAU) and *Yemeni* (YEM) dialects, and *Egyptian* (EGY) and *Sudanese* (SUD) dialects, respectively, exhibit unique patterns that may reflect historical and sociolinguistic factors.

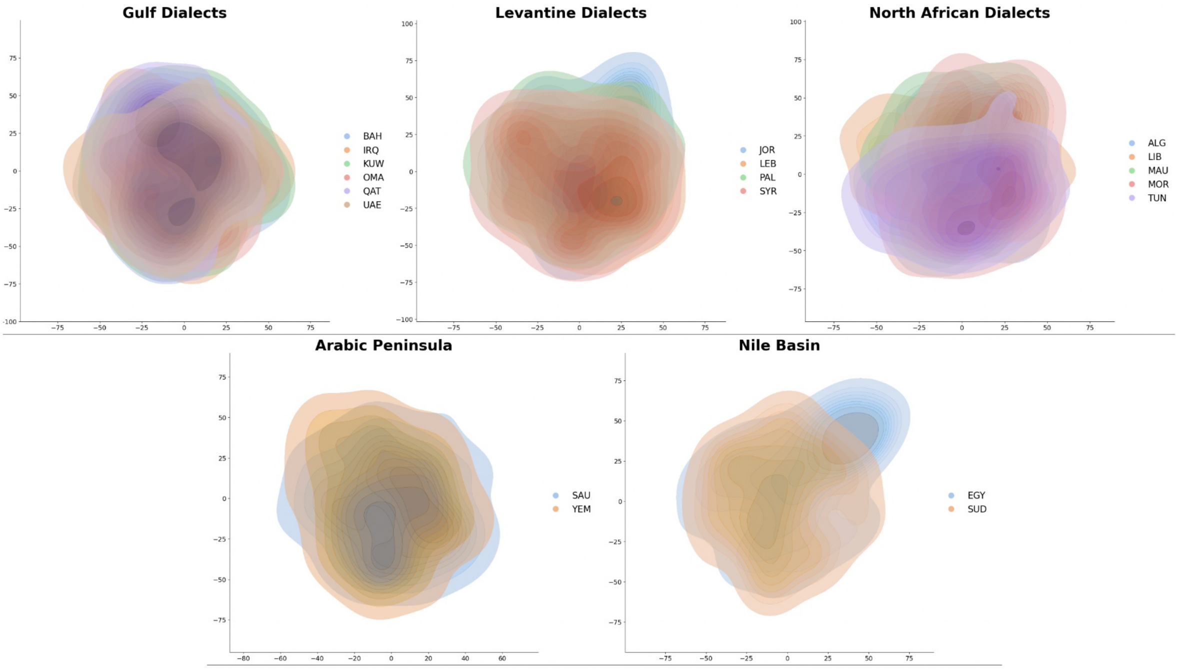


Figure 4: KDE plots of proverb embeddings from various Arabic dialects, highlighting both shared cultural themes and regional variations among Gulf, Levantine, North African, Arabian Peninsula and Nile Basin dialects.

3.2 Task Representation

Each proverb in *Jawaher* is represented with ***(a)*** an idiomatic, literal English translation, or English meaning, ***(b)*** an explanation in Arabic, and ***(c)*** an explanation in English. The explanations include stories detailing the themes and contexts in which the proverb is commonly used. Explanations, as such, enhance understanding and provide deeper insights into the meanings and cultural significance of proverbs.

To ensure high data quality, entries underwent dual annotation, consensus resolution of discrepancies, and pilot testing to identify and correct any annotation issues. More details regarding the quality assurance of data are provided in Appendix G.

Translation of Proverbs. In *Jawaher*, we provide translations of a total of 4,334 Arabic proverbs, offering different types of translations. All proverbs and their explanations were manually translated by a professional translator to ensure high quality and accurate representation of their meanings. Subsequently, two additional professional translators reviewed and revised the translations to verify their correctness and ensure they were correctly expressed idiomatically. We provide three types of translations of the proverbs: First, we provide (1) *English equivalents* for $1,289$ Arabic proverbs—these are similar idiomatic and figurative expressions in English. Since proverbs are inherently idiomatic, their translations should use common phrases and expressions of the target language (Baker, 2018; Gorjian, 2024), even if that means diverging from a literal translation to better capture the original's figurative or stylistic essence. Then, we offer (2) 840 *English meanings* of proverbs by paraphrasing them when *English equivalent* translations were not available. Finally, we provide (3) $2,205$ literal translations. Idiomatic expressions like proverbs are notoriously difficult to translate accurately (Gorjian, 2024). When direct equivalents in the target language are unavailable, translators resorted to providing literal translations. We believe that the inclusion of literal translations is essential in our work as it allows understanding the original wording and cultural nuances of the proverbs, which may not be fully captured through idiomatic or equivalent translations alone.

Arabic Explanation. Our dataset includes $3,764$ Arabic explanations that cover the stories behind the proverbs, how the proverbs are used in certain dialects, and the different situations in which these proverbs may be employed. Furthermore, for certain dialects such as *Lebanese*, and *Omani*, explanations include meanings of unusual words to provide more information for the reader to understand what these words mean in MSA.

English Explanation. *Jawaher* includes 2,500 human-translated English explanations of Arabic proverbs. These explanations cover proverbs with historical backgrounds or cultural stories from the Arab world, and we aim to have these explanations accurately convey the meanings behind the proverbs.

Full statistics for each task and examples across all dialects are provided in the Appendix C.

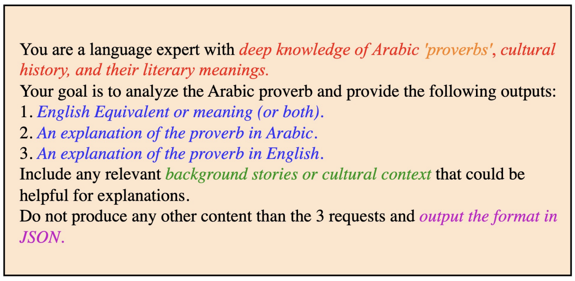
4 Experimental Setup

We evaluate *Jawaher* using both *open*- and *closed*-source state-of-the-art multilingual LLMs (mLLMs) to assess their abilities across our proposed tasks. The models are tested in a zero-shot setting (Sanh et al., 2021), allowing us to evaluate their inherent capacity to interpret, explain, and contextualize Arabic proverbs. To achieve this, we create a universal prompt template in English: (1) We set the *role* of the model as a language expert with deep knowledge of Arabic proverbs, cultural history, and literary meanings. (2) We deliver the *test input* to the model to produce the output. (3) We provide the name of the *task* that needs to be performed. (4) We specify the *context*, asking to include any relevant background stories or cultural context that could be helpful for explanations on the tasks. (5) Finally, we define what should be the expected *outcome* of the model. The prompt used to evaluate mLLMs can be found in Figure 5.

4.1 Models

Open Source. For *open*- source mLLMs, we evaluate two models from the Llama 3 family (Dubey et al., 2024): `Llama-3.1-8B-Instruct`, `Llama-3.2-3B-Instruct`, and `Google's Gemma-2-9B-it` (Team et al., 2024). These models were chosen for their varying sizes and multilingual capabilities, including support for Arabic.

Closed Source. We experiment with the following *closed*- source leading mLLMs: `GPT-4o` (Achiam et al., 2023), `Gemini 1.5 Pro` (Reid et al., 2024), `Claude 3.5 Sonnet`,[2] and `Cohere Command R+`.[3]



You are a language expert with *deep knowledge of Arabic 'proverbs', cultural history, and their literary meanings.*
Your goal is to analyze the Arabic proverb and provide the following outputs:
1. *English Equivalent or meaning (or both).*
2. *An explanation of the proverb in Arabic.*
3. *An explanation of the proverb in English.*
Include any relevant *background stories or cultural context* that could be helpful for explanations.
Do not produce any other content than the 3 requests and *output the format in JSON.*

Figure 5: Designed prompt used to test model's performance on *Jawaher*.

4.2 Evaluation setup

To evaluate the aforementioned models, we sample ten examples from each dialect and prompt the model using our crafted prompt. Below, we describe the automatic metrics and criteria we employ for human evaluation in order to assess the model's capabilities to understand and interpret the figurative language of Arabic proverbs across multiple dialects and cultural contexts.

Automatic Evaluation. We use BLEURT (Sellam et al., 2020) and BERTScore (Zhang et al., 2019) to judge the quality of explanations and translations. These metrics compute token similarity using contextual embeddings instead of exact matches, making them appropriate for measuring *fluency* and conveying the reference's *meaning*. Since BLEURT is calculated at the sentence level, we average these scores over the entire test set and report the final score on a scale between 0 and 100. BLEURT also only supports English; therefore, we only evaluate English-related tasks on BLEURT. For BERTScore, we report only the F_1 score. Full scores can be found in Appendix E.

Human Evaluation. Along with automatic evaluation, we conduct human evaluation to assess model performance. We develop evaluation criteria tailored to the nature of our tasks. Human evaluation is carried out by two expert annotators, both holding degrees in linguistics and translation studies, to ensure the accuracy and consistency of the evaluation criteria. For the first task, *translation*, we assess the models based on *accuracy* and *idiomaticity*, as the task inherently involves figurative language. Our goal is to determine how well the models understand the figurative aspects of the proverbs, translate them correctly, and convey their intended meanings. For the second task, providing *explanations* in both Arabic and English, we evaluate model output across four metrics: *clar-*

[2] https://www.anthropic.com/claude-3-5-sonnet

[3] `https://docs.cohere.com/docs/command-r-plus`

Source	Models	En. Trans.		En. Exp.		Ar. Exp.
		F_1	BLEURT	F_1	BLEURT	F_1
Open	Llama 3.1	89.42	30.12	89.19	46.51	67.05
	Llama-3.2	89.48	27.56	89.23	43.16	66.10
	Gemma	90.74	33.37	90.64	44.50	67.22
Closed	GPT-4o	**91.22**	44.50	**94.17**	40.68	**69.21**
	Gemini	89.73	38.53	89.69	**49.77**	68.62
	Comm. R+	90.09	39.91	89.93	45.36	68.34
	Claude	90.64	**44.70**	83.62	48.95	69.06

Table 1: BERTScore, reported as F_1, and BLEURT scores for english translation and english explanation. Scores represent the highest values in each source category, with the highest across all source categories highlighted in **bold**. Gemma: Gemma-2-9B-instruct. Gemini: Gemini 1.5 Pro. Claude: Claude 3.5 Sonnet.

ity, *depth & detail*, *correctness*, and *cultural relevance and sensitivity*. Each annotator evaluated the models' performance using the developed metrics, by scoring each response on a scale from 1 to 5. Appendix F outlines our evaluation metrics in more detail. **LLM as Judge.** Recently, LLM-based evaluation (Zeng et al., 2023; Chern et al., 2024) has emerged as a scalable and cost-effective alternative to human evaluations. In this work, we use *LangChain's*[4] evaluation framework with *string evaluator*. We specifically customize the evaluation criteria to our own and use `GPT-4o` for evaluation.

5 Results & Discussion

Table 1 reports BERTScores for all three tasks and BLEURT scores for *English translation* and *explanation* tasks.

5.1 Automatic Evaluation

Among open-source models, `Gemma2-9B-it` consistently outperforms both Llama models across all tasks in both metrics. It achieves the highest BLEURT scores in *English translation* (33.37) and *English explanation* (44.50) tasks. This aligns with the BERTScore results, where it shows the best performance among open-source models.

For closed-source models, `Claude 3.5 Sonnet` demonstrates the highest BLEURT score for *English translation* (44.70), while `Gemini 1.5 Pro` leads in *English explanation* (49.77). This contrasts with the BERTScore results, where `GPT-4o` shows superior performance across all tasks. Notably, `GPT-4o` also remains strong on BLEURT for English Translation (44.50), though it is slightly behind `Claude 3.5 Sonnet` on that metric.

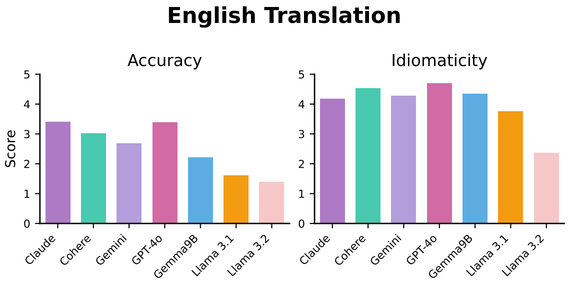


Figure 6: Results of human evaluation for both open- and closed-source models on two criteria in the English Translation task.

Looking across tasks, BLEURT scores for Arabic explanation reveal a noticeable performance drop for all models, suggesting that Arabic figurative language poses a greater challenge than English. While closed-source models generally outperform open-source ones, the performance differences vary by metric and task. This variability underscores the limitations of automatic evaluation when gauging how well models handle figurative language across different languages and prompts.

5.2 Human Evaluation

Two annotators were recruited for human evaluation, and we used *LangChain's* evaluation framework to simulate human annotation as described above. IAA using Krippendorff's α (Krippendorff, 2011) shows high agreement between the two human annotators, but a significant drop in correlation when including the third annotator (GPT-4o). Similarly, there is a notable drop in agreement scores for both the *English and Arabic explanation* tasks, with drops to 0.39 and 0.20, respectively. Consequently, we exclude model evaluations and report the average scores between the two annotators. The full IAA score is in Table 2.

Translation Quality. Figure 6 shows *accuracy* and *idiomaticity* scores for the *translation* task. Among closed-source models, `GPT-4o` achieves the highest score of (3.04) on *accuracy*. For open-source models, `Gemma2-9B-it` scores highest with 2.06. The Llama models perform the worst, with `Llama 3.1` and `Llama 3.2` scoring (1.55) and (1.36), respectively. This underperformance highlights a significant limitation in these models when producing nuanced linguistic and cultural content.

Idiomaticity scores, highlight models' ability to

[4] https://python.langchain.com/v0.1/docs

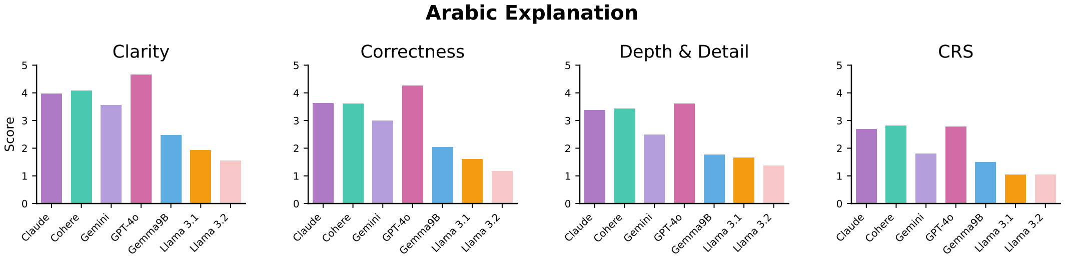


Figure 7: Results of human evaluation for both open- and closed-source models on four criteria in the Arabic explanation task. **CRS**: cultural relevance and sensitivity.

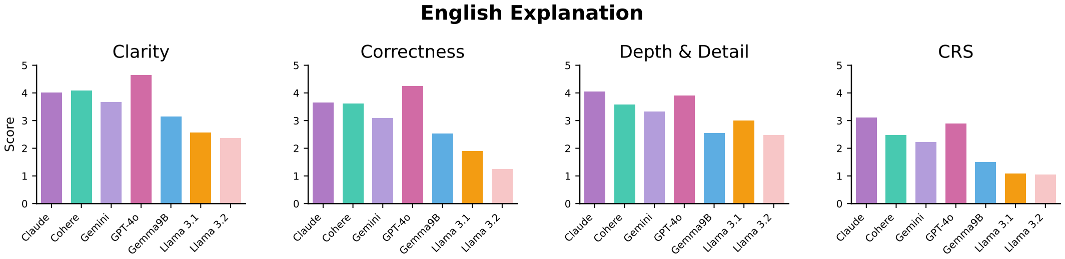


Figure 8: Results of human evaluation for both open- and closed-source models on four criteria in the English explanation task. **CRS**: cultural relevance and sensitivity.

Criteria	Human-only	Human-LLM
Accuracy	0.934	0.750
Idiomaticity	0.957	0.195
Clarity (Ar. exp.)	0.701	0.381
Correctness (Ar. exp.)	0.966	0.351
Depth and Detail (Ar. exp.)	0.976	0.547
Cultural rel. & Sensitivity (Ar. exp.)	0.703	0.050
Clarity (En. exp.)	0.155	0.289
Correctness (En. exp.)	0.949	0.678
Depth and Detail (En. exp.)	0.400	0.536
Cultural rel. & Sensitivity (En. exp.)	0.773	0.405

Table 2: Krippendorff's inter-annotator scores for different criteria across annotator groups for human-only agreements and human-LLM combined agreement.

produce natural and fluent translations. GPT-4o scores highest among all models tested with a score of (4.81). Other closed-source models also score high, with Cohere Command R+, Claude 3.5 Sonnet, and Gemini 1.5 Pro scoring (4.66), (4.59), and (4.58) respectively. Noticeably Gemma2-9B-it scores closely with closed-source models on this criterion, achieving (4.44). However, Llama models continue to show lower performance, scoring lowest among all models. Generally, both open- and closed-source models score higher on *idiomaticity*, indicating their ability to produce idiomatically correct translations. Regarding open-source models, we find that they often produce idiomatic expressions that are incorrect and fail to align with the meaning of the proverbs.

Arabic Explanation Quality. Figure 7 presents the results for the *Arabic explanation* task. GPT-4o scores highest in *clarity* (4.65), *correctness* (4.26), and *depth & details* (3.61), demonstrating its ability to deliver clear, accurate, and comprehensive explanations with underlying meaning and cultural nuances. Gemini 1.5 Pro scores lowest among closed- models, particularly in *cultural relevance and sensitivity* (1.80). Open-source models perform poorly across all criteria, with Gemma2-9B-it achieving the highest score among them at (2.47) in *clarity*, while the Llama models consistently score below 2 in all criteria.

Overall, closed-source models, particularly GPT-4o, outperformed across all *Arabic explanation* task metrics, excelling in generating clear, correct, and detailed explanations with cultural sensitivity. Although these models scores relatively high on *clarity* and *correctness*, their lower relative scores in *depth & detail* and *cultural relevance and sensitivity* imply these models are less capable of providing nuanced and context-rich explanations required to fully understand the proverbs. Open-source models showed significantly lower performance in explaining the stories behind the proverbs, their usage in certain dialects, and the situations in which these proverbs are used. Their explanations are generally short and lack cultural background stories.

English Explanation Quality. Figure 8 presents the results for the *English explanation* task. GPT-4o continues to score highest in *clarity* (4.65) and *correctness* (4.25), similar to its results in the *Arabic explanation* task, followed by Cohere Command R+ and Claude 3.5 Sonnet. For *depth & details*, Claude 3.5 Sonnet led with (4.04), and it also scores highest in *cultural relevance and sensitivity* (3.11). Open-source models, including Gemma2-9B-it and the Llama family models, perform significantly lower across all metrics, with Gemma2-9B-it achieving the best score among them in *clarity* (3.15).

Overall, closed-source models, particularly GPT-4o and Claude 3.5 Sonnet, outperform across all metrics, while open-source models lag, especially in *cultural relevance and sensitivity* and *depth & detail* in *explanation* tasks, indicating areas for improvement. English explanations are more detailed, with some reflecting cultural relevance, but none include the stories behind the proverbs. Results from both *English and Arabic explanation* tasks imply that models struggle to provide the nuanced and context-rich explanations required for fully understanding the proverbs in both languages.

6 Conclusion

We introduced *Jawaher*, a new benchmark for figurative language understanding in Arabic proverbs. Our manually curated dataset features proverbs from various Arabic dialects, paired with idiomatic translations and explanations, facilitating the development and evaluation of models capable of translating and explaining figurative language. Our experiments show that while both open- and closed-source mLLMs are able to provide idiomatically correct translations, they struggle with the deeper challenge of comprehending figurative language, particularly when generating detailed explanations that capture the stories behind the proverbs, dialect-specific nuances, and historical or cultural contexts discussed in our dataset. These findings highlight the need for continuous model improvements and dataset enrichment to address the complexities of fully understanding proverbs. We hope our work will inspire further interest in figurative language understanding, particularly across diverse Arabic dialects.

7 Limitations

We identify the following limitations in this work:

- This study primarily evaluates models under a zero-shot setting, which provides insights into their inherent abilities to understand and process Arabic proverbs without additional training. However, this approach may not reflect their full potential, particularly when it comes to more complex figurative language tasks. Future work could explore fine-tuning these models on proverb-specific datasets to enhance their performance.

- The human evaluation process, particularly in the context of cultural nuances, can be subjective and influenced by the evaluators' individual biases. This is especially relevant when assessing culturally grounded tasks like proverb explanation, where personal interpretation of cultural and regional references may affect the consistency and reliability of judgments.

- While the models used in this study claim to include Arabic in their training data, they are not fully optimized for Arabic or its dialectal variations. This limitation may have adversely impacted performance, especially when dealing with proverbs that rely heavily on unique linguistic and cultural nuances found across different Arabic-speaking regions.

Ethics Statement

Promoting Cultural Sensitivity and Representation in AI. Proverbs are deeply embedded in cultural contexts, epitomizing the wisdom, values, and traditions of a community. Recognizing the cultural significance of Arabic proverbs, our research prioritizes the accurate representation and interpretation of these expressions within LLMs. *Jawaher* includes proverbs from various Arabic dialects, ensuring a broad and inclusive representation of the linguistic diversity inherent in the Arabic-speaking world. By addressing Anglo-centric biases prevalent in many LLMs, our work promotes cultural sensitivity and strives to create AI systems that are more equitable and reflective of diverse cultural narratives. This focus on cultural representation is important for developing AI tools that are both effective and respectful of the rich cultural heritage they aim to serve.

Advancing the Understanding of Figurative Language in AI. Proverbs represent a form of figurative language that poses unique challenges for automated interpretation and generation by LLMs. Our evaluation of *Jawaher* highlights the current limitations of LLMs in capturing the cultural nuances and contextual relevance inherent in proverbs. This focus on figurative language advances the technical capabilities of AI systems and contributes to a deeper understanding of how language models can better align with human cultural and linguistic diversity.

Data Privacy. *Jawaher* dataset comprises proverbs that are publicly available and do not contain any personal or sensitive information. By utilizing data from public sources, we eliminate privacy concerns and ensure compliance with ethical guidelines and institutional policies regarding data usage.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[5] and UBC Advanced Research Computing-Sockeye.[6]

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Salwa Ahmed. 2005. *Educational and social values expressed by proverbs in two cultures: knowledge and use of proverbs in Sudan and in England*. Ph.D. thesis, Berlin, Techn. Univ., Diss., 2005.

Juhaina Maen Al Issawi. 2021. Cultural-bound meaning of animal names in arabic. *English Linguistics Research*, 10(1):42–55.

Manar M Almanea. 2021. Automatic methods and neural networks in arabic texts diacritization: a comprehensive survey. *IEEE Access*, 9:145012–145032.

Walid Aransa. 2015. *Statistical machine translation of the Arabic language*. Ph.D. thesis, Université du Maine.

Muhammad Hasan Bakalla. 2023. *Arabic culture: through its language and literature*. Taylor & Francis.

Mona Baker. 2018. *In other words: A coursebook on translation*. Routledge.

Riadh Belkebir and Nizar Habash. 2021. Automatic error type annotation for arabic. In *Conference on Computational Natural Language Learning*.

Hezi Brosh. 2013. Proverbs in the arabic language classroom. *International Journal of Humanities and Social Science*, 3(5):19–29.

Lucília Chacoto. 1994. *Estudo e formalização das propriedades léxico-sintácticas das expressões fixas proverbiais*.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing simile knowledge from pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5875–5887, Dublin, Ireland. Association for Computational Linguistics.

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*.

[5] https://alliancecan.ca

[6] https://arc.ubc.ca/ubc-arc-sockeye

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Ashutosh Dwivedi, Pradhyumna Lavania, and Ashutosh Modi. 2023. EtiCor: Corpus for analyzing LLMs for etiquettes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6921–6931, Singapore. Association for Computational Linguistics.

Tayyara A El-Rahman. 2020. The practicability of proverbs in teaching arabic language and culture. *Language Teaching Research*.

NOUH SABRI Elmitwally and Ahmed Alsayat. 2020. Classification and construction of arabic corpus: figurative and literal. *Journal of Theoretical and Applied Information Technology*, 98(19).

Mohammed Farghal. 2021. Animal proverbs in jordanian popular culture: A thematic and translational analysis. *Journal of English Literature and Language*, 2(1):1–8.

Charles A Ferguson. 2003. Diglossia. In *The bilingualism reader*, pages 71–86. Routledge.

Susan Fussell and Mallie Moss. 2008. Figurative language in emotional communication.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Raymond W Gibbs and Herbert L Colston. 2012. *Interpreting figurative meaning*. Cambridge University Press.

Mª González Rey. 2002. Isabel, la phraséologie du français.

Bahman Gorjian. 2024. Translating english proverbs into persian: A case of comparative linguistics. *Comparative Linguistics*. Ph.D. dissertation.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2023. Investigating the applicability of self-assessment tests for personality measurement of large language models. *arXiv preprint arXiv:2309.08163*.

Hady J Hamdan, Hanan Al-Madanat, and Wael Hamdan. 2023. Connotations of animal metaphors in the jordanian context. *Psycholinguistics*, 33(1):132–166.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7875–7887, Dublin, Ireland. Association for Computational Linguistics.

Clive Holes. 2004. *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, et al. 2024. Emma-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint arXiv:2409.17892*.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *arXiv preprint arXiv:2203.07785*.

Anubha Kabra, Emmy Liu, Simran Khanuja, Alham Fikri Aji, Genta Winata, Samuel Cahyawijaya, Anuoluwapo Aremu, Perez Ogayo, and Graham Neubig. 2023. Multi-lingual and multi-cultural figurative language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8269–8284, Toronto, Canada. Association for Computational Linguistics.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Nathalie Aussenac-Gilles, and Lamia Hadrich Belguith. 2015. Towards a contextual pragmatic model to detect irony in tweets. In *53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, volume 2, pages 644–650. ACL: Association for Computational Linguistics.

Grgur Kovač, Masataka Sawayama, Rémy Portelas, Cédric Colas, Peter Ford Dominey, and Pierre-Yves

Oudeyer. 2023. Large language models as superpositions of cultural perspectives. *arXiv preprint arXiv:2307.07870*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Technical report, University of Pennsylvania ScholarlyCommons.

EV Kuhareva. 2008. Arabic proverbs and sayings. *Dictionary with lexical and phraseological comments. Moscow: AST: Vosto. Zapad*.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2023. Are multilingual llms culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. *arXiv preprint arXiv:2309.08591*.

W. Mieder. 2004. *Proverbs: A Handbook*. Greenwood Folklore Handbooks. Bloomsbury Academic.

Wolfgang Mieder. 2007. *Proverbs as cultural units or items of folklore*. na.

Wolfgang Mieder. 2021. From innovative anti-proverbs to modern proverbs. *The Discoursal Use of Phraseological Units*, page 61.

Suzanne Mpouli. 2017. Annotating similes in literary texts. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2014. Brighter than gold: Figurative language in user generated comparisons. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 2008–2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ben Prystawski, Paul Thibodeau, Christopher Potts, and Noah D Goodman. 2022. Psychologically-informed chain-of-thought prompts for metaphor understanding in large language models. *arXiv preprint arXiv:2209.08141*.

Zhaozhi Qian, Faroq Altam, Muhammad Saleh Saeed Alqurishi, and Riad Souissi. 2024. Cameleval: Advancing culturally aligned arabic language models and benchmarks. *arXiv preprint arXiv:2409.12623*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Aida Rinasovna Fattakhova, Alfiya Albertovna Gimadeeva, Emma Rishatovna Galimova, and Timur Akzamovich Shaihullin. 2019. Proverbs in arabic: Definition, classification, outlook reflection. *Journal of Research in Applied Linguistics*, 10(Proceedings of the 6th International Conference on Applied Linguistics Issues (ALI 2019) July 19-20, 2019, Saint Petersburg, Russia):521–528.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-flute: Visual figurative language understanding with textual explanations. *arXiv preprint arXiv:2405.01474*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ekaterina Shutova. 2011. Computational approaches to figurative language.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419.

Pushpdeep Singh, Mayur Patidar, and Lovekesh Vig. 2024. Translating across cultures: Llms for intralingual cultural adaptation. *arXiv preprint arXiv:2406.14504*.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? a probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407, Held Online. INCOMA Ltd.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Faezdzhon Meliboevich Tursunov. 2022. The influence of the cultural aspect on the translation of proverbs and idioms (a case study of the tajiki/persian, russian and english languages). *Filologicheskie nauki. Voprosy teorii i praktiki*, 15(2):616–620.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Kees Versteegh. 2014. *Arabic language*. Edinburgh University Press.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Shelia Webster. 2000. Arabic proverbs and related forms. *De Proverbio: An Electronic Journal of International Proverb Studies*, 6(2).

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Jiali Zeng, Linfeng Song, Jinsong Su, Jun Xie, Wei Song, and Jiebo Luo. 2020. Neural simile recognition with cyclic multitask learning and local attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9515–9522.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

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

Appendices

We offer an addition structured as follows:

A Development of Figurative Language.

Prior work on figurative language understanding has covered a wide range of topics, including simile detection and generation (Niculae and Danescu-Niculescu-Mizil, 2014; Mpouli, 2017; Zeng et al., 2020). Although language models can often recognize non-literal language and may attend to it less, they still struggle to fully capture the implied meaning of such phrases (Shwartz and Dagan, 2019). More recent studies have shifted toward tasks focused on *comprehending* figurative language (Chakrabarty et al., 2022; He et al., 2022; Prystawski et al., 2022; Jang et al., 2023). Several studies have explored whether knowledge of figurative meaning is encoded in the learned representations by investigating how well these models capture non-literal meanings (Tan and Jiang, 2021; Chen et al., 2022; Dankers et al., 2022; Jang et al., 2023). Additionally, efforts in dataset development have focused on more diverse tasks aimed at comprehending figurative language, such as recognizing textual entailment (RTE) (Chakrabarty et al., 2022; Kabra et al., 2023), multilingual understanding(Liu et al., 2023), and tasks involving figurative language beyond text (Yosef et al., 2023; Chakrabarty et al., 2023; Saakyan et al., 2024). However, the lack of comprehensive datasets focusing on figurative language, particularly Arabic proverbs, limits the ability to thoroughly evaluate LLMs' cultural awareness and their true understanding of culturally embedded non-literal expressions.

B Linguistic Background of Arabic

Arabic is spoken by approximately 450 million people worldwide and is the sole or joint official language in over twenty Middle Eastern and African nations, including Morocco, Algeria, Mauritania, Tunisia, Libya, Egypt, and Sudan. MSA is a descendant of Classical Arabic (CA), which was used in the 6th century and is the language of the Quran. While MSA has evolved from CA, it has been significantly simplified to suit contemporary literary, poetic, and official discourse. Although its syntactic structure remains unchanged, MSA continues to evolve in vocabulary and phraseology. This pan-Arab variety is widely used in written forms and media, including news broadcasts, political speeches, and public ceremonies. Arabic is a highly sophisticated and intricate language, characterized by its rich morphological and grammatical features (Holes, 2004). The complexity of the Arabic language stems from its orthographic, morphological, semantic, and syntactic systems (Aransa, 2015).
Moreover, the Arabic writing system uses diacritics, which are small marks placed above or below letters. These marks, called "vocalization," help with the correct pronunciation and meanings of words. They are important for clearing up many ambiguities in the text, ensuring accurate understanding and interpretation (Almanea, 2021). For example, the orthographic ambiguity of the diacritic in the word كتب without diacritics can be ambiguous. However, when diacritics are added, it can take on different meanings: كَتَبَ (kataba): With a fatha on each consonant is a verb that means "he wrote." كُتُب(kutub): With a damma on the first and second consonants is a noun which means "books." كُتِبَ(kutiba): With a damma on the first consonant and a kasra on the second consonant is a verb in the passive voice, and means "it was written." These diacritics provide clarity and precision and ensure that the intended meaning is conveyed accurately. The complexity and diversity of Arabic create significant challenges for NLP. This is due to the language's rich morphology, diverse dialects, and intricate script, which includes diacritics (Abdul-Mageed et al., 2020; Belkebir and Habash, 2021). These factors make tasks such as text processing, speech recognition, and machine translation more difficult compared to other languages. Understanding and addressing these challenges is crucial for improving NLP applica-

tions for Arabic speakers. Arabic language exhibits considerable linguistic diversity, which includes numerous dialects spread across a vast geographical area (Versteegh, 2014) and is recognized as a diglossic language as it exhibits two varieties (MSA and various dialects) which are used interchangeably (Ferguson, 2003). Proverbs, as succinct and poignant expressions, reflect the values, wisdom, and social norms prevalent within various Arabic-speaking communities. By including proverbs from a wide range of Arabic dialects, it increases the complexity of the Arabic language. They serve as a tool for understanding the historical, social, and cultural contexts from which they originate.

C Data Statistics.

Table 7 shows full statistics of our dataset. Table 5 shows examples of all proposed tasks.

D Themes of Arabic proverbs

Arabic uniquely maintains a well-defined link between past and present, classical and modern, in structure, grammar, vocabulary, and cultural traditions (Versteegh, 2014). This linguistic and cultural continuity is best represented in proverbs, which are deeply rooted in Arab culture and identity (El-Rahman, 2020) and convey popular wisdom across generations (Rinasovna Fattakhova et al., 2019). Understanding them requires grasping the underlying cultural values and social norms they reflect, as their rich semantic layers cannot be derived from mere literal interpretation (Ahmed, 2005). The animal names take over most of the Arabic proverbs (Al Issawi, 2021), For instance, *lion* symbolizes strength, courage, royalty, and ferocity in Arab culture. It has multiple figurative meanings that are widely used in the proverbs such as هذا الشبل من ذاك الأسد ("this little lion from that lion") which highlights the similarity between a father and his son, usually in a positive light. The proverb is used to emphasize the inherited qualities or behaviors that are apparent between generations. Also,the lion often signifies a large portion, as in the proverb حصة الأسد ("the lion's share"), indicating that a person has received the most significant portion.

Another common theme in Arabic proverbs is the use of *camels*, underscoring the Bedouin lifestyle's deep connection with this animal. Camels are integral to the daily lives of Bedouins and farmers, particularly in arid and semi-arid regions (Al Issawi, 2021). They are used for riding during peace and war, transporting burdens, and providing essential resources like milk and meat.

By comparing proverbs across the countries we investigate, we noticed that camel is a common theme and is used throughout proverbs with similar meanings, but with different words. For example, in Libya, they say الجملْ ما يِحقّ عوجْ رقْبْتهْ, while in Levantine, they use لو كان الجمل بيشُوفْ حردبنه كان بيقع بيفك رقبته. Both proverbs are used in describing a person who sees the flaws of others but does not see their own flaws. It is noticeable that there is a significant similarity in the proverbs across different Arab countries, whether in the East or the West, both in their wording and their meaning. This similarity stems from the strong cultural connections and the ongoing communication between these countries. More examples of the camel theme are listed in table 3.

Some Arabic proverbs originate from true stories that occurred in the past. These proverbs have been passed down through generations and are still widely used today. These historical anecdotes provide context and meaning to the proverbs, highlighting the practical lessons and experiences of earlier times. For instance, the story behind a well-known MSA proverb على نفسها جَنَتْ براقش. This proverb dates back to the Pre-Islamic era. Baraqaš was a dog who inadvertently led enemy soldiers to her people through her barking. Thus, it is said, "Baraqaš brought it upon her people" or "She brought it upon herself." This ancient proverb is used to describe someone who harms themselves by their own actions; an illustrated example is provided in figure 9.

E Automatic Evaluation

Full results including Precision and Recall scores can be found in Table **??**.

F Human Evaluation Metrics

In our evaluation of LLMs on the *Jawaher* dataset, we assess them on two main tasks: translation and providing explanations of proverbs across 20 Arabic varieties in both English and Arabic.

For the **translation task**, we evaluate the models' ability to provide the English equivalents of the proverbs. To measure the capabilities of these models to understand figurative language such as proverbs and to provide idiomatic translations, we assess the models on two main scales: *accuracy* and *idiomaticity*, each rated on a scale from 1 to 5. A summary of the metric in table 4

Accuracy ensures that the meaning is correct and how well the translation conveys the exact intent and cultural context with precise language, ensuring complete accuracy and clarity.

5: The translation fully captures the original proverb's meaning, intent, and cultural context with precise language, ensuring complete accuracy and clarity.

4: The translation is mostly accurate, capturing the original meaning and cultural context well, but it may miss some subtle nuances or minor details without significantly altering the overall meaning.

3: The translation conveys the general meaning of the proverb but lacks some nuance or cultural context, resulting in minor inaccuracies or partial misinterpretations.

2: The translation captures some elements of the original meaning, but it lacks important details or cultural context, leading to noticeable inaccuracies or a partial misinterpretation of the proverb.

1: The translation fails to convey the correct meaning or completely misrepresents the original proverb's intent and cultural context, or the translation is missing altogether.

The other criteria we assess is the idiomaticity of the translated proverbs.

Idiomaticity refers to how naturally the translation fits into the linguistic and cultural norms of the target language. A proverb should sound like something a native speaker would say, using phrases, metaphors, or expressions that are common in the target language. An idiomatic translation may diverge from a literal translation to better match the figurative or stylistic nature of the original proverb. Since proverbs are idiomatic expressions, their translations must feel natural in the target language. The translation should sound like a proverb in the target language rather than a forced, awkward sentence.

5: The translation is completely natural, idiomatic, and would be readily recognized as a proverb by native speakers.

4: The translation is almost natural and idiomatic, with only slight room for improvement.

3: The translation is generally understandable but contains notable awkwardness or unnatural phrasing.

2: The translation is somewhat understandable but remains somewhat awkward and far from idiomatic.

1: The translation is completely awkward or unnatural and is not idiomatic in the target language.

The second task in our evaluation is the **explanation** of the proverbs in both Arabic and English. We evaluated the models' outputs across four metrics: *clarity*, *correctness*, *depth and detail*, and *cultural relevance and sensitivity*.

Clarity: By *clarity* we assess how clearly the meaning of the proverb is explained. The explanation should be understandable, and the evaluator should not need further clarification.

5: The explanation is perfectly clear and easy to understand, with no confusion.

4: The explanation is mostly clear and easy to follow, with only minor areas needing clarification.

3: The explanation is somewhat clear, but there are one or two areas that are confusing or lack detail.

2: The explanation is somewhat unclear, with several confusing or ambiguous points.

1: The explanation is very unclear or confusing, making it difficult to understand the meaning of the proverb.

Correctness: *Correctness* criterion evaluates the accuracy of the explanation. It should reflect the actual meaning of the proverb without misinterpretation.

5: The explanation is completely correct and accurately conveys the meaning of the proverb.

4: The explanation is mostly correct, with only minor inaccuracies.

3: The explanation is partially correct but misses some key aspects or provides an incomplete interpretation.

2: The explanation is mostly incorrect, with significant inaccuracies or misinterpretations.

1: The explanation is incorrect, misinterpreting the proverb's meaning.

Depth and Detail *Depth and details* criterion measures how comprehensive the explanation is. It should cover any underlying meanings, cultural nuances, and possible interpretations, offering more than just a surface-level description.

5: The explanation is highly detailed, providing a thorough understanding of the proverb's meaning, cultural context, and potential interpretations.

4: The explanation is fairly detailed, covering cultural nuances or different interpretations with

only minor omissions.
3: The explanation gives a basic understanding but does not delve deeply into cultural nuances or multiple interpretations.
2: The explanation provides some details but lacks sufficient depth or misses key elements.
1: The explanation is overly simplistic, lacking depth and failing to cover important aspects of the proverb.

Cultural Relevance and Sensitivity This criterion assesses whether the explanation acknowledges any cultural context necessary to fully understand the proverb. Proverb explanations should consider whether the audience understands the cultural or historical background if needed.
5: The explanation fully considers cultural context, making it highly relevant and meaningful to the target audience.
4: The explanation is culturally relevant, addressing most of the important cultural or contextual elements.
3: The explanation somewhat considers the cultural context but could do more to connect with the audience's cultural understanding.
2: The explanation makes only minimal reference to cultural relevance, leaving important aspects unaddressed.
1: The explanation does not account for the cultural context or gives an inappropriate explanation that does not resonate with the target audience.

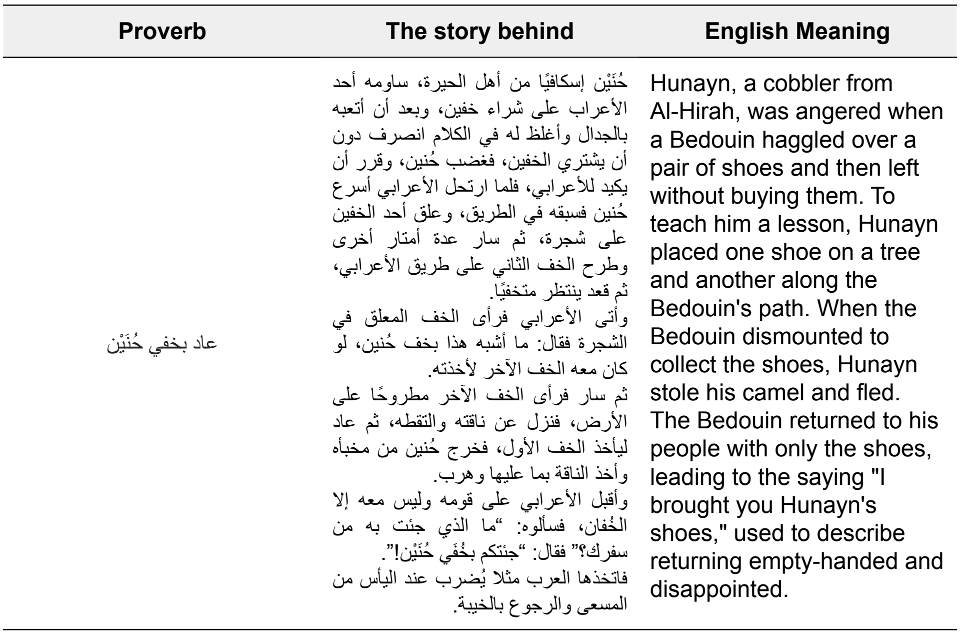
Table 4 outlines the full criteria of the evaluation metrics.

G Quality Assurance of the Data

To ensure the high quality of our dataset and the subsequent evaluation process, we implemented a dual annotation procedure. Three native-speaking expert linguists meticulously verified the data by first reviewing the information gathered from online sources. They checked for language accuracy, clarity of explanations, and corrected any grammatical errors or other issues present in the data. For the translation tasks, dual annotation was employed to assess the quality of translations comprehensively. This involved reviewing the English equivalents, meanings, and explanations to ensure consistency and accuracy across all tasks. After completing these quality assurance steps, we proceeded with our evaluation.

Variety	Proverb	English Translation	Shared Meaning
Algeria	اللّي تلمّه النّملة في عام يأكله الجمل في لُقمة	What an ant collects in a year, a *camel* eats in one bite.	These proverbs emphasize the idea that small, long -term efforts or savings can easily be taken or consumed by something much larger in an instant.
Egypt	اللّي تجمعه النّملة في سنة يأخده الجمل في خُفّه	What an ant gathers in a year, a *camel* takes in its hoof.	
Iraq	اللّي تجمعه النّملة بسنة يشيله البعير بخُفّه	What an ant gathers in a year, a *camel* carries in its hoof.	
Libya	اللّي تحوّشه النّملة في عامْ يأخذها الجمل في خُفّه	What an ant saves in a year, a *camel* takes in its hoof.	
Qatar	يجمعه العصفور في سنة ويأكله الجمل في لقمة	What a bird gathers in a year, a *camel* eats in one bite.	
Tunisia	اللّي تلمّو النّملة في عام يهزّو الجمل الجمل في فم	What an ant collects in a year, a *camel* sways in its mouth.	

Table 3: Proverbs about *camels* from six Arab countries illustrating a shared concept that long-term savings can easily be consumed or lost in an instant.



Proverb	The story behind	English Meaning
عاد بخفي حُنَيْن	حُنَيْن إسكافيًا من أهل الحيرة، ساومه أحد الأعراب على شراء خفين، وبعد أن أتعبه بالجدال وأغلظ له في الكلام انصرف دون أن يشتري الخفين، فغضب حُنين، وقرر أن يكيد للأعرابي، فلما ارتحل الأعرابي أسرع حُنين فسبقه في الطريق، وعلق أحد الخفين على شجرة، ثم سار عدة أمتار أخرى وطرح الخف الثاني على طريق الأعرابي، ثم قعد ينتظر متخفيًا. وأتى الأعرابي فرأى الخف المعلق في الشجرة فقال: ما أشبه هذا بخف حُنين، لو كان معه الخف الآخر لأخذته. ثم سار فرأى الخف الآخر مطروحًا على الأرض، فنزل عن ناقته والتقطه، ثم عاد ليأخذ الخف الأول، فخرج حُنين من مخبأه وأخذ الناقة بما عليها وهرب. وأقبل الأعرابي على قومه وليس معه إلا الخُفان، فسألوه: "ما الذي جئت به من سفرك؟" فقال: "جئتكم بخُفَي حُنَيْن!". فاتخذها العرب مثلا يُضرب عند اليأس من المسعى والرجوع بالخيبة.	Hunayn, a cobbler from Al-Hirah, was angered when a Bedouin haggled over a pair of shoes and then left without buying them. To teach him a lesson, Hunayn placed one shoe on a tree and another along the Bedouin's path. When the Bedouin dismounted to collect the shoes, Hunayn stole his camel and fled. The Bedouin returned to his people with only the shoes, leading to the saying "I brought you Hunayn's shoes," used to describe returning empty-handed and disappointed.

Figure 9: An MSA example illustrating the historical story behind a well known proverb.

Task	Measure	Scale
Translation	Accuracy	**5**: Fully captures meaning, intent, and cultural context precisely. **4**: Mostly accurate; minor nuances missed without altering overall meaning. **3**: Conveys general meaning but lacks some nuances or context. **2**: Captures some elements but lacks important details or context. **1**: Fails to convey correct meaning or misrepresents intent; may be missing.
	Idiomaticity	**5**: Completely natural and idiomatic; recognized by native speakers. **4**: Almost natural; slight improvement possible. **3**: Understandable but contains awkward phrasing. **2**: Somewhat understandable but remains awkward. **1**: Completely awkward; not idiomatic.
Explanation	Clarity	**5**: Perfectly clear and easy to understand. **4**: Mostly clear; minor clarification needed. **3**: Somewhat clear; some confusing areas. **2**: Unclear; several confusing points. **1**: Very unclear; difficult to understand.
	Correctness	**5**: Completely correct; accurately conveys meaning. **4**: Mostly correct; minor inaccuracies. **3**: Partially correct; misses key aspects. **2**: Mostly incorrect; significant inaccuracies. **1**: Incorrect; misinterprets meaning.
	Depth and Detail	**5**: Highly detailed; thorough understanding. **4**: Fairly detailed; minor omissions. **3**: Basic understanding; lacks depth. **2**: Lacks depth; misses key elements. **1**: Overly simplistic; lacks important aspects.
	Cultural Relevance and Sensitivity	**5**: Fully considers cultural context; highly relevant. **4**: Culturally relevant; addresses most elements. **3**: Somewhat considers cultural context; could connect better. **2**: Minimal cultural relevance; important aspects unaddressed. **1**: Does not account for cultural context; inappropriate explanation.

Table 4: A summary of the evaluation metrics for translation and explanation tasks of our work

Variety	Example	En Equivalent	Ar. Explanation	En. Explanation
ALG	حوحو يشكر روحو	Don't brag about yourself let others praise you	يقال لمن يمدح نفسه ويشكرها تنمرا عليه حوحو يشكر روحو	Used to mock someone who praises themselves excessively.
BHR	إنفخ يا شريم قال ماكو برطم	You can not escape bad luck	الشخص الذي لا يفهم ما تقول له، او تشرح له.	When someone is trying to explain something to a person who can't understand.
EGY	آخرةِ خدمة الغُز علقة	The end of favor is denial.	يقال للدلالة على نكران الجميل، ومقابلة الإحسان بالشر	Used to describe ingratitude, where good deeds are met with harm or betrayal.
IRQ	يتعلّم الحجامة بروس اليتامة	A burnt child dreads the fire.	الشخص الذي يستغل الاشخاص الفقراء، لكي يقوم بعدة تجارب فاشلة	Refers to someone experimenting recklessly at the expense of the weak or poor.
JOR	أتبدلت غزلانها بقرودها	A falling master makes a standing servant	يضرب عند تغير الأحوال وحلول الرديء مكان الجيد	Describes a situation where good circumstances are replaced by bad ones.
KWT	حاط دِوبَه دوبي	He picks on me	جعلني محط اهتمامه في الشجار والمجادلة	Describes someone who targets another person in arguments or disputes.
LBN	الي بكبِّر لقمتو، بِغصّ فَيا.	Don't live beyond your means	يضرب في التّحذير من تخطي الاعتدال.	This proverb urges moderation.
LIB	اضرب القطوس، تخاف العروس	beat the dog before the lion	أي اضرب الضعيف، ترهب القوي	Strike the weak, intimidate the strong.
MRT	الدني ماجاهَ حواش	Do not spare anything in this life	من ينفق من ما اعطاه الله بيسر وبدونِ تقتير	Encourages living freely and without excessive restraint.
MOR	اش عرف الحمير في سكينجبير	Casting pearls before swine	جهل قيمة الاشياء الثمينة التي لا يقدرها حق قدره	When valuable things are wasted on those who cannot appreciate them.
MSA	تجري الرياح بما لا تشتّهي السفن	Things don't always go as they're planned.	أذّ الحياة لا تسير دائما وفق رغبات الإنسان	Life doesn't always unfold according to one's desires or plans.
OMA	برمة الشرك لا تثور	Two cooks spoil the cook	يقصد به المتزوجين الذين يريدون إدارة المنزل في نفس الوقت	Describes chaos that arises when too many people try to take control.
PAL	أَعمِص وبِيِتجعمص	A beggar with a havana	الشخص الذى يتعالى على الناس وهو ليس مثلهم	Someone who is acting superior but doesn't actually belong to that status.
QAT	البيض اللى وقع على بيض فقشه	The rotten apple injures its neighbors	يقصد به الضعيف تظهر حقيقته اذا قابل ضعيفاً مثله	Weakness is exposed when confronted by another weak entity.
SAU	ما عنده إلا الخرطي	Be all talk and no action	الشخص الثرثار الذي لا يعمل ويقول ما لا يفعل	Someone who talks a lot but has no action behind their words.
SUD	أبيض جناح اسود مراح	Fine feathers do not make fine birds	يصف من يرتدي فاخر الثياب البيضاء، وتكون داره مسودة من القذارة	Describes someone who looks good on the outside but their true nature is the opposite.
SYR	ابن الديب ما بيتربى	The apple doesn't fall far from the tree	يقال لمِن يحاول أن يغيِّر خصالاً أصيلة في إنسان ما	Describes how people often inherit their parent's traits, good or bad.
TUN	الجمل مايراش كربته يرا كربة غيرو	One does not see one's own defects.	يقصد بها أن الشخص لا يرى عيوبه لكن يرى عيوب غيره	People who are quick to point out others' faults while ignoring their own.
UAE	اصبوعي في حلوجهم وصبوعهم في عيوني	Evil in return for good deed	عن نبذ الجحود، من يفعل الخير الكثير و لا يلقى سوى النكران	When good deeds are repaid with ungratefulness or evil actions.
YEM	تزين بالخلاخل، والبلا من داخل	Beauty is only skin deep	الرجل الذي يختار أن يتزوج امرأة جميلة، ولكنه لا ينظر إلى أخلاقها	Refers to a man who decides to marry a beautiful woman, but he does not care about her morals.

Table 5: Examples from 20 Arabic varieties arranged in alphabetical order include the variety, Arabic proverbs, English equivalents, Arabic explanations, and English explanations

Source	Models	English Equivalent			English Explanation			Arabic Explanation		
		P	**R**	**$F_{1.0}$**	**P**	**R**	**$F_{1.0}$**	**P**	**R**	**$F_{1.0}$**
Open	Llama 3.1	87.28	91.67	89.42	86.83	91.68	89.19	67.31	66.94	67.05
	Llama-3.2	87.41	91.66	89.48	86.92	91.66	89.23	65.88	66.50	66.10
	Gemma9B-instruct	88.81	92.76	90.74	88.53	92.85	90.64	67.23	67.35	67.22
Closed	GPT-4o	89.59	92.91	91.22	92.42	95.98	94.17	67.33	71.29	69.21
	Gemini 1.5 Pro	87.89	91.66	89.73	87.59	91.91	89.69	66.75	70.69	68.62
	Cohere Command R+	88.05	92.24	90.09	87.65	92.33	89.93	65.97	70.99	68.34
	Claude 3.5 Sonnet	90.53	90.76	90.64	81.34	86.07	83.62	67.27	71.04	69.06

Table 6: BERTScore table. T1: English Equivalent, T2: English Explanation, T3: Arabic Explanation

Variety	No. Proverbs	Ar Explanation	En Explanation	En Equivalent
ALG	312	153	108	54
BAH	134	112	103	25
EGY	1,018	781	224	314
IRQ	304	104	104	26
JOR	398	132	127	26
KUW	126	102	102	44
LEB	495	257	111	24
LIB	390	114	109	40
MAU	360	127	112	32
MOR	507	255	159	77
MSA	604	228	199	185
OMA	182	100	100	30
PAL	1,078	210	121	60
QAT	161	152	139	33
SAU	302	197	111	51
SUD	228	111	102	46
SYR	1,723	199	102	45
TUN	1,221	168	126	57
UAE	212	137	135	71
YEM	282	125	106	49
Total	10,037	3,764	2,500	1,289

Table 7: Summary of data statistics for 20 Arabic varieties, listed alphabetically. The table includes the total *No.: Number of Arabic proverbs, their *Ar: Arabic explanations, *En: English explanations, and *En: English equivalents.