Beyond Literal Meaning: How LLMs Interpret Yemeni Proverbs

Nasser Thmer[1,2] , Ali Al-Laith[3], Muhammad Shoaib[1]
[1]Computer Science Department,University of Engineering and Technology Lahore, Pakistan,
[2]Department of Computer, College of Education Lawder, University of Abyan , Abyan, Yemen,
[3] Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark
nasserthmer.net@gmail.com, alal@di.ku.dk, shoaib@uet.edu.pk

Abstract

We present a benchmark Yemeni proverbs dataset paired with expert-annotated explanations, designed to evaluate the cultural reasoning abilities of large language models (LLMs). Using zero-shot and few-shot prompting, we assess seven LLMs through both automatic and human evaluation. Results show that instruction-tuned models like GPT-4o and Gemini 1.5 Pro outperform smaller models in both automatic and human evaluations. Few-shot prompting significantly improves performance across all models, underscoring its value for figurative and culturally grounded language tasks. Notably, ALLaM, a bilingual model trained on Arabic and English, achieves competitive results, demonstrating the potential of regionally adapted models for low-resource cultural tasks. LLM-as-a-Judge evaluation correlates strongly with human assessment (Kendall's up to 0.98). Error analysis identifies recurring literal interpretation and cultural misalignment as key failure modes.

Keywords: Yemeni Proverbs, Figurative Language, Cultural NLP, Benchmark Dataset, Large Language Models

1. Introduction

Proverbs encapsulate cultural wisdom, offering insights into a society's values, beliefs, and linguistic creativity. However, their figurative and context-dependent nature makes them challenging for natural language processing (NLP) systems to interpret. While prior work has explored proverb processing in widely studied languages, low-resource languages, particularly Arabic dialects like Yemeni, remain understudied. Yemeni proverbs, rich in metaphorical and region-specific expressions, present a unique test case for evaluating the cultural and linguistic adaptability of modern large language models (LLMs).

Recent advances in LLMs have demonstrated remarkable capabilities in understanding figurative language(Liu et al., 2022; Al-Laith et al., 2025). Yet, their performance on culturally nuanced tasks, especially in low-resource settings, is still limited by dependencies on training data and prompting strategies. While zero-shot and few-shot learning reduce reliance on fine-tuning, their effectiveness for proverbs requiring deep cultural context remains unclear. Similarly, fine-tuning offers potential for specialization but demands curated datasets, which are scarce for dialects like Yemeni Arabic.

In this work, we address key gaps in culturally-aware NLP by introducing a novel dataset of Yemeni proverbs paired with expert-annotated explanations, providing a new benchmark for evaluating figurative and culture-specific language understanding. We systematically assess the performance of LLMs under zero-shot and few-shot prompting to evaluate their ability to interpret these proverbs. Additionally, we analyze model successes and failures to highlight core challenges in processing culturally grounded and metaphorical expressions.

Our comprehensive analysis not only benchmarks model performance but also validates LLM-as-a-Judge as a reliable evaluation method, showing a strong correlation with human assessment (Kendall's τ of up to 0.976). Furthermore, we identify systematic error patterns—including literal interpretation, cultural misalignment, and explanation ambiguity—that persist across models, providing insights into the fundamental challenges of processing culturally grounded figurative language. These findings underscore the need for both improved evaluation frameworks and enhanced cultural reasoning capabilities in multilingual models. The code and dataset are available in this repository: [1].

2. Related Work

Understanding and explaining culturally grounded proverbs, such as those in the Yemeni dialect, poses unique challenges for NLP models, especially in capturing figurative meaning and contextual nuance.

The ePiC dataset evaluates language models' ability to interpret proverbs in context, requiring analogical reasoning beyond surface-level comprehension(Ghosh and Srivastava, 2021). Tasks such as recommendation, narrative generation, and motif identification expose performance gaps between neural models and humans, emphasizing the complexity of figurative language. Figurative expres-

[1] https://github.com/NasserThmer/Yemeni-Proverbs-dataset-code

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1071–1080
11–16 May 2026.

sions also pose challenges for machine translation, as models often fail to capture non-compositional meanings without cultural and contextual grounding(Gupta and Poonia, 2014). ERP studies reveal that literal meanings integrate more easily than figurative ones, highlighting cognitive demands in proverb interpretation(Ferretti et al., 2020). Kemper (1981) further shows that extended context improves figurative understanding, though word-level cues can bias interpretation. Overall, while benchmarks like ePiC offer evaluation frameworks, effective proverb processing still requires models capable of cultural and contextual reasoning.

Several studies have developed proverb corpora via ethnographic and textual analysis. For example, Migdadi et al. (2023) compiled 5,634 Jordanian Arabic proverbs from natural discourse and media, thematically categorized but lacking large-scale machine-readable annotations. Cross-linguistic work compares Arabic proverbs to those in English, Danish, Persian, and other languages, revealing universal themes and culture-specific features through qualitative methods such as interviews and manual annotation(Alharbi, 2024; Ababneh, 2025; Zuheiri and Tai, 2023). Translation studies note frequent semantic loss when rendering Arabic proverbs into English or French, advocating paraphrasing or cultural equivalents as alternatives(Alfaleh, 2020; Jabak, 2022; Hmaidan, 2024; Aminou, 2023; Pedersen et al., 2025). Evaluations of machine translation systems (e.g., Bing, Google) confirm over-reliance on literal translation, while communicative strategies improve outcomes but require deep cultural understanding(Jibreel, 2023). Jawaher, a benchmark for evaluating models' ability to interpret Arabic proverbs across dialects, covered few samples in Yemeni dialects(Magdy et al., 2025). Despite growing interest, computational resources for Arabic proverbs, especially for underrepresented dialects like Yemeni, remain limited.

3. Dataset Description

3.1. Collection and Challenges

The dataset was compiled from both printed and online sources focused on Yemeni proverbs. Printed materials included *Dictionary of Common Proverbs: Collection, Preparation, and Study*(Al-Hamdani, 2013), *The Yemeni Wealth of Popular Proverbs: Collection and Explanation*(Al-Adimi, 1987), *The Common Proverbs of Yafa'*(Al-Khalaqi, 2013), and *Qutoof Min Al-Amthal Al-Yemeniya*(Salem, 2024). Proverbs from Al-Adimi (1987) were manually transcribed, while the others were digitized using OCR, followed by manual verification due to the challenges posed by the complexity of the Arabic script and poor scan quality.

Two online sources were also used: the Old Sanaa website[2] and Folk Memory from Folk Culture Magazine[3]. Data extraction was performed via web scraping, with prior permission from site owners to ensure ethical compliance.

The collection process faced several challenges. Limited digital resources on the Yemeni dialect required significant manual effort. OCR inaccuracies and inconsistent or poorly structured explanations necessitated careful post-editing to produce a clean, coherent dataset. All explanations were preserved or lightly rephrased for clarity without altering their original meaning, ensuring the dataset maintains linguistic and cultural authenticity.

3.2. Statistics

The final dataset consists of 456 instances, each comprising a short figurative expression (typically a proverb) paired with a human-authored explanation. All examples were manually curated to ensure high quality and consistency. As shown in Table 1, the dataset exhibits notable variability in both proverb and explanation length. On average, explanations are significantly longer than the source texts, reflecting their role in elaborating implicit meanings. These structural characteristics underscore the interpretive nature of the task and suggest that computational models must handle varying input lengths and generate contextually rich outputs.

Table 1: Descriptive statistics of the dataset.

Statistic	Value
Total Samples	456
Total Words (Texts + Explanations)	9,855
Unique Words (Total)	3,709
Average Words per Proverb	5.22
Average Words per Explanation	16.39
Max Proverb Length (words)	20
Min Proverb Length (words)	2
Max Explanation Length (words)	29
Min Explanation Length (words)	7

4. Methodology

4.1. Models

We evaluate diverse LLMs for proverb interpretation, including: (1) closed-source models (GPT-4o(OpenAI, 2024), Gemini 1.5 Pro) for multilingual baselines; (2) Arabic-optimized models (Jais-13B(Elmadany et al., 2023),

[2] https://old-sanaa.org/

[3] https://folkculturebh.org/ar/index.php?issue=67&page=article&id=500

ALLaM-7B(Bari et al., 2025)) for cultural alignment; and (3) general-purpose models (LLaMA-3-8B(Meta AI, 2024), Mistral-7B(Jiang et al., 2023), DeepSeek-7B(DeepSeek-AI, 2024)) as architecture controls. This selection enables comparisons across training paradigms and cultural/linguistic capabilities in zero-shot and few-shot settings.

4.2. Prompts

The following prompts were used consistently across all experiments to elicit model-generated explanations of Yemeni proverbs, serving as the basis for both zero-shot and few-shot evaluations.

Zero-Shot

Zero-Shot Prompt (English Instruction)

System Prompt:
You are a linguistic expert in Yemeni proverbs. Given the proverb below, explain its figurative meaning clearly and concisely in Modern Standard Arabic.

Input Format:
`{proverb}`

Few-Shot

System Prompt:
You are a linguistic expert in Yemeni proverbs. Given the proverb below, explain its figurative meaning clearly and concisely in Modern Standard Arabic.

Example 1
Proverb:
الصبر لا يشبع الجيعان، ولا يكسي العريان

Translation: Patience doesn't put food on the table or clothes on your back.
Explanation:
يعني أن الصبر وحده لا يكفي لحل المشكلات المادية، مثل الجوع أو الفقر، بل لا بد من العمل أو المساعدة.

Translation: Patience alone won't fix material hardship; you need action or support to solve problems like hunger or poverty.

Example 2
Proverb:
القبر تزله فيه، وعاده ما يكفيه

Translation: Give him the world and he'd still want more. (or: "Greed has no bottom.")
Explanation:
يعبر عن الجشع والطمع الذي لا يتنهي. يشير إلى أن بعض الناس لا يرضون حتى بعد أن يأخذوا كل شيء.

Translation: It portrays relentless greed: some people are never satisfied, no matter how much they take—even to the point that not even the grave would "be enough."
Target Format:

Proverb: {proverb}
Explanation:

4.3. Evaluation Metrics

We evaluate seven language models on Yemeni proverbs using automatic metrics over the full dataset and human evaluation on a consistent 5% random sample across models and settings , under both zero-shot and few-shot conditions.

Automatic Evaluation: We use three automatic metrics: Cosine Similarity, BERTScore(Zhang et al., 2019), and Semantic Answer Similarity (SAS)(Risch et al., 2022), using the Arabic-SBERT-100K sentence transformer model(Reimers and Gurevych, 2019). Cosine Similarity and BERTScore capture semantic similarity using sentence embeddings, while SAS employs a fine-tuned cross-encoder to assess the semantic equivalence between a generated explanation and a human reference. This makes SAS particularly suitable for evaluating open-ended generation tasks such as proverb interpretation, where multiple semantically valid explanations may exist. These metrics are especially relevant in this context, as semantic equivalence is more important than surface-level lexical similarity.

Human Evaluation: Following Magdy et al. (2025), we conduct a reference-based human evaluation to assess the quality of the generated explanations using three criteria: accuracy, cultural appropriateness, and clarity. Three annotators compared each explanation to a gold-standard reference created or verified by native speakers and rated each criterion on a 5-point Likert scale(van der Lee et al., 2019). The annotators were selected from distinct Yemeni dialect regions to ensure cultural grounding. Accuracy measured how well the explanation conveyed the intended meaning of the proverb; cultural appropriateness assessed alignment with Yemeni values and traditions; and clarity evaluated the explanation's readability and coherence. Annotators were also encouraged to provide optional comments to support their ratings. Krippendorff's inter-annotator agreement (IAA)(Krippendorff, 2011) demonstrates a high level of consistency among the three annotators from different Yemeni dialect backgrounds

To complement human evaluation and automatic metrics, we experimented with using LLMs as evaluators. The motivation is that while human judgments are reliable, they are resource-intensive, and automatic metrics often fail to capture figurative and cultural nuances. We employed gpt-oss-120b as the evaluation model, prompting it with explicit instructions to act as a linguistic expert in Yemeni proverbs and to score generated explanations on *Accuracy*, *Cultural Appropriateness*, and *Clarity* using a 5-point Likert scale. The evaluation prompts were designed to mirror the human annotation guidelines, and deterministic decoding parameters (temperature = 0, top-p = 1) were used to ensure reproducibility.

4.4. Human Evaluation Guidelines

We outline the human evaluation protocol used to assess the quality of machine-generated explanations for Yemeni proverbs. Annotators were presented with three elements: the proverb, a reference (human-written) explanation, and the explanation generated by the model. They were asked to evaluate the generated explanation based on the following three core criteria:

- **Accuracy:** Does the explanation correctly capture the intended figurative meaning of the proverb?
- **Cultural Appropriateness:** Does the explanation reflect Yemeni cultural norms and contextually appropriate interpretations?
- **Clarity:** Is the explanation coherent, well-structured, and easy to understand?

Each criterion was rated independently on a five-point Likert scale ranging from 1 (very poor) to 5 (excellent). This structured evaluation framework ensures consistency across annotators and enables reliable human-centered assessment of explanation quality.

4.5. Experimental Setup

We conduct generation experiments on 456 proverbs from our dataset to (1) establish a baseline for zero-shot and few-shot performance, (2) evaluate the generalization capacity of Arabic-compatible LLMs, and (3) analyze their ability to interpret culturally rich, dialect-specific expressions without prior domain adaptation.

In the zero-shot setting, models were prompted with a simple instruction to explain a given proverb clearly and concisely, without any fine-tuning or in-context examples. This setup tests how well models generalize to unfamiliar, culturally specific tasks(Brown et al., 2020). In the few-shot setting, we appended two manually selected proverb–explanation pairs to the prompt from the training set. These exemplars captured a range of cultural themes and linguistic variations. This setup evaluates the model's ability to adapt to new tasks with minimal guidance(Brown et al., 2020).

5. Results and Analysis

5.1. Automatic Evaluation

Automatic evaluation (Figure 1) using Cosine Similarity, BERTScore, and Semantic Answer Similarity showed that GPT-4o and ALLaM-7B performed best overall, with Gemini 1.5 Pro excelling in semantic similarity (81.65%) in the few-shot setting. Few-shot prompting significantly improved performance across models, particularly for DeepSeek and Jais, highlighting the benefit of exemplar guidance. While BERTScore tended to be high across models, Semantic Answer Similarity revealed large performance gaps, especially for smaller models like Mistral-7B, underscoring challenges in capturing figurative and culturally grounded meanings.

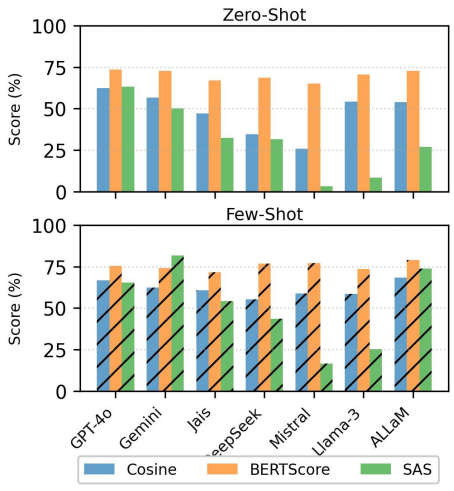

Figure 1: Automatic evaluation results across models.

Table 2 provides a detailed comparison across all evaluated models.

5.2. Human Evaluation

Human evaluation (Figure 2) further confirmed the superiority of GPT-4o, Gemini 1.5 Pro, and ALLaM-7B in accuracy, cultural appropriateness, and clarity. Few-shot prompting enhanced clarity across all models, though smaller or open models like Mistral-7B, DeepSeek, and Jais remained limited in producing culturally appropriate and semantically accurate explanations. Overall, results indi-

Table 2: Automatic evaluation results using Cosine Similarity, BERTScore, and Semantic Answer Similarity (SAS) under Zero-Shot and Few-Shot settings.

Model	Zero-Shot			Few-Shot		
	Cosine	BERTScore	SAS	Cosine	BERTScore	SAS
GPT-4o	**62.38%**	**73.57%**	**63.29%**	66.72%	75.34%	65.30%
Gemini 1.5 Pro	56.76%	72.58%	50.16%	62.47%	73.97%	**81.65%**
Jais-Adapted-13B-Chat	47.02%	67.09%	32.55%	60.60%	71.49%	54.12%
DeepSeek-LLM-7B-Chat	34.54%	68.54%	31.68%	55.17%	76.74%	43.65%
Mistral-7B-Instruct-v0.2	26.01%	64.95%	03.27%	58.93%	77.11%	16.73%
Meta-Llama-3-8B-Instruct	54.23%	70.50%	08.49%	58.55%	73.62%	25.46%
ALLaM-7B-Instruct-Preview	54.04%	72.72%	27.01%	**68.29%**	**78.91%**	73.92%

cate that large instruction-tuned models better handle abstract, culturally specific tasks, while few-shot examples offer a valuable mechanism for improving interpretability and output quality in low-resource cultural contexts.

Table 3 reports the mean and standard deviation (SD) scores under both Zero-Shot and Few-Shot prompting conditions.

Overall, GPT-4o and Gemini 1.5 Pro achieved the highest scores across all criteria, especially in clarity (up to **4.54 ± 0.59**), with consistent performance gains under Few-Shot prompting. Among open models, ALLaM-7B showed relatively strong results (e.g., Clarity = **3.58 ± 0.69**), outperforming other open-source models such as Jais, DeepSeek, and Mistral, which struggled in both accuracy and cultural alignment. Few-Shot prompting improved performance across most models, highlighting the effectiveness of in-context examples in enhancing the cultural and explanatory quality of generated outputs.

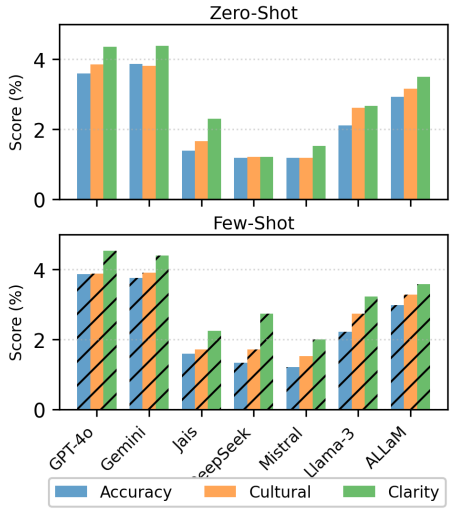

Figure 2: Human evaluation results across models.

5.3. LLM as a Judge

To validate the reliability of LLM-as-a-Judge, we compared its scores against human annotations on the same sample subset, reporting Spearman correlations between 0.6–0.8 depending on the criterion. We observed that the LLM aligned most strongly with human judgments of *Clarity*, while it was less consistent on *Cultural Appropriateness*, reflecting its limited cultural grounding. We emphasize that LLM-based evaluation is complementary to human annotation and should be interpreted with caution, given potential biases in the evaluation model.

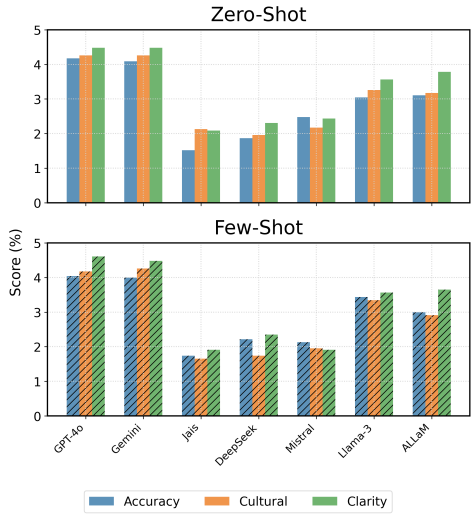

Figure 3: LLM-as-a-Judge evaluation results in Zero-Shot and Few-Shot settings.

6. Analysis

6.1. Metric Reliability

Human vs. LLM-as-Judge Agreement. We evaluate the agreement between human evaluation and LLM-as-Judge using Kendall's rank correlation τ with two-sided significance tests. As shown

Table 3: Human evaluation results for Accuracy, Cultural Appropriateness, and Clarity under Zero-Shot and Few-Shot settings. Scores are reported as Mean and Standard Deviation (SD).

Model	Zero-Shot						Few-Shot					
	Accuracy		Cultural Approp.		Clarity		Accuracy		Cultural Approp.		Clarity	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
GPT-4o	3.594	0.864	3.855	0.784	4.362	0.643	3.870	0.827	3.884	0.769	4.536	0.592
Gemini 1.5 Pro	3.870	1.104	3.812	0.724	4.377	0.831	3.754	0.939	3.913	0.906	4.391	0.715
Jais-Adapted-13B-Chat	1.391	0.851	1.667	0.853	2.304	1.570	1.594	0.921	1.710	1.228	2.246	1.495
DeepSeek-LLM-7B-Chat	1.188	0.374	1.217	0.384	1.217	0.397	1.333	0.471	1.710	0.544	2.739	0.696
Mistral-7B-Instruct-v0.2	1.188	0.576	1.188	0.576	1.522	0.618	1.217	0.397	1.522	0.511	2.000	0.778
Meta-Llama-3-8B-Instruct	2.116	1.242	2.609	0.821	2.667	1.385	2.217	0.832	2.739	0.835	3.232	0.775
ALLaM-7B-Instruct-Preview	2.928	1.218	3.159	1.063	3.493	1.044	2.986	0.728	3.275	0.743	3.580	0.691

Table 4: Kendall's τ (Human vs. LLM-as-Judge) with two-sided p-values for Zero-Shot and Few-Shot.

2*Metric	Zero-Shot		Few-Shot	
	τ	p-value	τ	p-value
Accuracy	0.683	0.0334	0.714	0.0302
Clarity	0.781	0.0151	0.976	0.0024
Cultural	0.683	0.0334	0.683	0.0334

in Table 4, all correlations are positive and statistically significant ($p < 0.05$), demonstrating a reliable alignment between automatic and human assessments. The Few-Shot setting achieves near-perfect agreement on *Clarity* ($\tau = 0.976$), while *Accuracy* and *Cultural* criteria maintain strong consistency across both settings ($\tau \approx 0.68$–0.71). These results validate LLM-as-Judge as a reliable evaluator that closely tracks human preferences, particularly for clarity assessment under Few-Shot prompting.

Weak Generated Explanations This subsection presents examples of weak explanations produced by various language models when interpreting Yemeni proverbs. These outputs commonly exhibit limitations such as literal interpretation, lack of cultural sensitivity, vague phrasing, or irrelevant content. Such weaknesses are particularly evident in smaller or non-instruction-tuned models, which often fail to capture the figurative and contextual meaning of the proverbs. Representative samples are provided in Table 5, illustrating common failure patterns observed across models and experimental setups.

Strong Generated Explanations In contrast, this subsection presents examples of strong generated explanations that accurately capture both the literal and figurative meanings of the proverbs while maintaining cultural relevance and clarity. These outputs are predominantly produced by larger, instruction-tuned models such as GPT-4o and Gemini 1.5 Pro, especially in the few-shot setting. Table 6 includes selected high-quality outputs demonstrating the models' potential for culturally sensitive language understanding.

6.2. Error Analysis

A qualitative error analysis was conducted to identify and categorize the recurring shortcomings observed in the generated explanations of Yemeni proverbs across the evaluated models. Several common error patterns emerged from this analysis.

1. **Ambiguity and Incompleteness:** Some explanations were vague, unclear, or truncated, failing to fully express the intended figurative meaning of the proverb.
2. **Irrelevance:** In several cases, models produced explanations that were off-topic or introduced content unrelated to the original proverb, indicating a lack of contextual grounding.
3. **Literal Interpretation:** Models often interpreted proverbs literally, overlooking their metaphorical or idiomatic meaning, which is crucial for accurate understanding.
4. **Cultural Misinterpretation:** Some outputs distorted or misrepresented culturally embedded concepts, revealing limitations in the models' cultural sensitivity and regional knowledge.

Table 7 presents representative examples of each error type, highlighting the challenges that current models face in processing figurative and culturally embedded language.

7. Discussion

Our findings indicate that although LLMs can produce lexically and contextually similar explanations, capturing the deeper semantic and cultural nuances of Yemeni proverbs remains challenging, particularly in zero-shot settings. Employing few-shot prompting significantly enhances performance, underscoring the importance of example-driven methods for culturally specific tasks. However, discrep-

Model	Proverb	Reference Explanation	Generated Explanation	Score
GPT-4o	كتن و فجي	يشير هذا المثل إلى الشعور بالظلم والقهر الذي قد يواجهه الإنسان في مواقف معينة.	المثل اليمني آكتن و فجي يعبر عن فكرة الصبر والانتظار للحصول على نتائج أفضل...	2.78
Gemini 1.5 Pro	كتن و فجي	يشير هذا المثل إلى الشعور بالظلم والقهر...	يشير إلى التناقض والاختلاف الشديد...	2.00
ALLaM-7B-Instruct-Preview	الغريق يتعلق بقشه	يشير هذا المثل إلى أن الشخص في حالة اليأس أو الشدة...	يبذل جهدًا كبيرًا في سبيل الحصول على شيء بسيط...	1.56
DeepSeek-LLM-7B-Chat	السراج المغنزي ولا الغدره	وجود شيء بسيط أفضل من لا شيء...	السراج الذي يحتوي على ماء ليس الغدره.	1.00
Meta-Llama-3-8B-Instruct	لو حلقوا لابن عمك بل لحيتك	ضرورة الاستعداد لمواجهة نفس المصير...	يهتم بشؤون الآخرين ويترك شأنه الخاص...	1.00
Jais-Adapted-13B-Chat	ابدا بيمينك ولو بيسرك علي	البدء بالأمور الصحيحة رغم الظروف...	مثل عمني قديم جدا.	1.00
Mistral-7B-Instruct-v0.2	ابدا بيمينك ولو بيسرك علي	البدء بالأمور الصحيحة رغم الظروف...	يعني أن تبدأ بالعمل ولو كان ع...	1.00

Table 5: Weak Generated Explanations for Selected Yemeni Proverbs Across Language Models

Model	Proverb	Reference Explanation	Generated Explanation	Score
GPT-4o	اللي ما يخاف من الله خاف منه	يشير هذا المثل إلى أن الشخص الذي لا يخشى الله في أفعاله قد يكون خطرًا على الآخرين.	المثل اليمني اللي ما يخاف من الله خاف منهٔ يحمل معنى تحذيريًا. يشير إلى أن الشخص الذي لا يخشى الله ولا يراعي القيم الدينية والأخلاقية قد يكون خطرًا وغير موثوق به...	5.0
Gemini 1.5 Pro	المجد ولا الغني	يشير هذا المثل إلى أن المجد والرفعة أفضل من الغنى الذي قد يؤدي إلى الطغيان والغرور.	يشير إلى أن الشرف والكرامة أهم من الثروة المادية...	5.0
ALLaM-7B-Instruct-Preview	اللي ما يخاف من الله خاف منه	يشير هذا المثل إلى أن الشخص الذي لا يخشى الله في أفعاله قد يكون خطرًا على الآخرين.	هذا المثل يعني أن من لا يخاف من الله يجب أن يخاف من الناس...	4.67
Meta-Llama-3-8B-Instruct	الناس علي دين ملوكهم	يشير هذا المثل إلى أن الشعوب تتأثر بسلوك قادتها...	الناس يتبعون سلوك وآراء زعمائهم وقادتهم...	4.56
Jais-Adapted-13B-Chat	اللي ما يخاف من الله خاف منه	يشير هذا المثل إلى أن الشخص الذي لا يخشى الله في أفعاله قد يكون خطرًا على الآخرين.	معناه أن الشخص الذي يخاف من الله لا يخاف من أحد.	4.33
Mistral-7B-Instruct-v0.2	الدهر قلبه يقلبه	يشير هذا المثل إلى أن الشخص قد يواجه أوقاتًا جيدة وأوقاتًا سيئة في حياته.	يعني أن الزمان يتقلب بسرعة وتتغير أحوال الناس فيه...	3.0
DeepSeek-LLM-7B-Chat	الدهر قلبه يقلبه	يشير هذا المثل إلى أن الشخص قد يواجه أوقاتًا جيدة وأوقاتًا سيئة في حياته.	المعنى مجازي له هو: يتحرك بسرعة لا يصدق ويضحك بصوت كبير.	1.67

Table 6: Strong Generated Explanations for Selected Yemeni Proverbs Across Language Models

ancies observed in metrics such as Semantic Answer Similarity highlight that high scores may not necessarily reflect genuine cultural comprehension, thereby emphasizing the necessity of supplementary human evaluations.

Overall, instruction-tuned models like GPT-4o and Gemini 1.5 Pro exhibit superior generalization, cultural sensitivity, and clarity. Furthermore, the ALLaM model series, trained on approximately 1.2 trillion mixed-language tokens (540 billion Arabic tokens—half of which are translated—and 660 billion English tokens), underscores the critical role of extensive and diverse training data in enabling effective figurative and culturally nuanced language generation. Conversely, smaller open models frequently generate literal or culturally mismatched explanations. These outcomes reinforce the significance of model scale, diverse training data, and targeted prompting techniques in facilitating effective figurative and cross-cultural language generation.

8. Conclusion and Future Work

This study introduced a benchmark dataset of Yemeni proverbs with expert-annotated explanations and used it to evaluate the interpretive abilities of leading LLMs. Through automatic and human evaluation, we showed that larger models outperform smaller ones, particularly in accuracy, clarity, and cultural appropriateness. Few-shot prompting proved especially effective in enhancing output quality.

Additionally, our analysis validates LLM-as-a-Judge as a reliable evaluation method for cultural NLP tasks, while revealing that literal interpretation and cultural misalignment remain fundamental challenges that require specialized architectural solutions beyond simple model scaling.

Future work will explore fine-tuning and retrieval-augmented generation to further improve cultural grounding. Additionally, we aim to develop more reliable automatic evaluation metrics that better reflect figurative and cultural understanding, advancing the evaluation of culturally informed NLP systems.

Proverb	Reference Explanation	Generated Explanation	Error Type
الحر تخفيه الإشارة والذل يقرح بالقضا	الإنسان الحُر يكفيه التلميح، بينما الذليل لا يستجيب إلا بالعقوبة القاسية.	ء المحكم على المسلمين في اليمن والتي تعتبر من…	Ambiguity & Incompleteness
الفريع له صميل	الشخص العنيد أو المتكبر يحتاج إلى من يردعه بالقوة أو العقوبة.	بن عمرو بن عبد العزيز بن عبد الله بن عبد الرحم…	Irrelevance
ثر الطيور من رجم مرتين	أسوأ الناس من لا يتعلم من أخطائه ويكررها.	.	Incomplete Explanation
اللي ما تعرف تخبز ترجع الخلل عند المافي	الشخص الذي لا يتقن عمله يلوم الأدوات أو الآخرين بدلاً من نفسه.	ا والسلحين والمستخدمين للقوى السياسية والاقتص…	Irrelevance & Ambiguity
ما حاج إلا من جع برجله	صاحب الحاجة هو من يقوم بنفسه بمتابعة وإنجاز حاجته.	للانتقال إلى المسجد الأقصى والمسجد النبوي والم…	Literal Misinterpretation
الشراع بشرع لعمري والحطبه بحطبه	كل شخص مسؤول عن تصرفاته وأفعاله ولا يتحمل مسؤولية غيره.	لعمري هي أمران تعتبر غير مسموح بها في الإسلام.…	Cultural Misinterpretation
ثربه بين كلاب	شيء ثمين وقع بين أيدي أشخاص لا يعرفون قيمته.	.	Incomplete Explanation

Table 7: Examples of error patterns identified in model-generated explanations of Yemeni proverbs.

9. Limitations

While our findings provide valuable insights into LLMs' performance on culturally grounded figurative language, several limitations remain. First, the dataset focuses solely on Yemeni Arabic proverbs, which, although rich and underrepresented, limit generalizability to other dialects and languages. Second, the number of few-shot exemplars is fixed and small, which may not fully capture the range of variation needed for optimal in-context learning. Additionally, while human evaluation provides nuanced feedback, it is resource-intensive and subject to annotator interpretation. Finally, models were evaluated without fine-tuning, meaning their performance reflects general pretrained capabilities rather than task-specific adaptation.

10. Ethics Statement

This study highlights the need for cultural sensitivity in AI systems that interpret Arabic proverbs, particularly from the Yemeni dialect. Since proverbs express deep social and moral wisdom, literal or culturally uninformed interpretations can distort their meaning. To prevent this, the researchers curated a linguistically authentic dataset of Yemeni proverbs with expert explanations.

The work underscores ethical and inclusive AI development, addressing the biases of English-centric models and the difficulties large language models face with dialectal and figurative language. Ultimately, it advocates for NLP systems that respect cultural context, not just achieve surface accuracy, ensuring that AI interpretations align with the communities they represent.

11. Bibliographical References

Sana' Ababneh. 2025. A cross-cultural study: Animal proverbs, the case of english and colloquial arabic. *Journal of Ecohumanism*.

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, Bolette Pedersen, Carsten Levisen, and Daniel Hershcovich. 2025. Dying or departing? euphemism detection for death discourse in historical texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1353–1364, Abu Dhabi, UAE. Association for Computational Linguistics.

Bodour Abdulaziz Alfaleh. 2020. Translation quality assessment of proverbs from english into arabic: The case study of one thousand and one english proverbs translated into arabic. *Arab World English Journal*.

Adel Alharbi. 2024. Conceptualization of pragmatic language through proverbs: A comparative study of arabic and english proverbs. *International Journal of Educational Sciences and Arts*.

Mohamadou Aminou. 2023. Adaptation in translation: A case study of transferring some french proverbs into arabic. *International Journal of Humanities and Educational Research*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Todd R Ferretti, A. Katz, Christopher A. Schwint, Courtney Patterson, and Dagna Pradzynski. 2020. How discourse constraints influence neurolinguistic mechanisms during the comprehension of proverbs. *Cognitive, Affective, & Behavioral Neuroscience*, 20:604 – 623.

Sayan Ghosh and Shashank Srivastava. 2021. epic: Employing proverbs in context as a benchmark for abstract language understanding. *ArXiv*, abs/2109.06838.

Ranu Gupta and Sandeep Kumar Poonia. 2014. Study of machine translation in nlp systems: Non compositional idioms and proverbs.

Marvet Abed Ahmad Hmaidan. 2024. Proverbs translation techniques from english into arabic among jordanian university students. *Journal of Ecohumanism*.

Omar Osman Jabak. 2022. Proposed taxonomy of strategies for translating english proverbs into arabic. *Journal of Translation and Language Studies*.

Dr. Ibrahim Jibreel. 2023. Online machine translation efficiency in translating fixed expressions between english and arabic (proverbs as a case-in-point). *Theory and Practice in Language Studies*.

S. Kemper. 1981. Comprehension and the interpretation of proverbs. *Journal of Psycholinguistic Research*, 10:179–198.

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

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Samar Mohamed Magdy, Sang Yun Kwon, Fakhraddin Alwajih, Safaa Taher Abdelfadil, Shady Shehata, and Muhammad Abdul-Mageed. 2025. JAWAHER: A multidialectal dataset of Arabic proverbs for LLM benchmarking. In *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, Albuquerque, New Mexico. Association for Computational Linguistics.

Hamzah Faleh Migdadi, Bara'ah Abdejmajeed Al-Ababneh, Khalid Alsmadi, Yasmeen Almadani, and Bowroj Sameh Taany. 2023. A corpus-based study of proverbs in colloquial jordanian arabic: A socio-pragmatic analysis. *Theory and Practice in Language Studies*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318. ACL.

Bolette S. Pedersen, Nathalie Sørensen, Sanni Nimb, Dorte Haltrup Hansen, Sussi Olsen, and Ali Al-Laith. 2025. Evaluating LLM-generated explanations of metaphors – a culture-sensitive study of Danish. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 470–479, Tallinn, Estonia. University of Tartu Library.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julian Risch, Marvin Schröder, Johannes Daxenberger, and Iryna Gurevych. 2022. Semantic answer similarity for evaluating question answering systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4784–4797.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan. Association for Computational Linguistics.

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

Nahi Abid Ibrahim Al Zuheiri and Mahdi Ahmed Hussein Al Tai. 2023. A general view on arabic and persian proverbs and the major approaches on their translation. *Journal of Language Studies*.

12. Language Resource References

n.d. *Yemeni Proverbs*. Old Sanaa. Accessed: 3 March 2025.

n.d. *Yemeni Proverbs in the Folk Memory*. Folk Culture. Accessed: 3 March 2025.

Mohammed Othman Thabet Al-Adimi. 1987. *The Yemeni Wealth of Popular Proverbs: Collection and Explanation*. Self-published.

Ahmed Ali Al-Hamdani. 2013. *Dictionary of Common Proverbs: Collection, Preparation, and Study*. University of Aden Publishing and Printing House.

Ali Saleh Al-Khalaqi. 2013. *The Common Proverbs of Yafa'*. Abadi Center for Studies and Publishing, 3.

M Saiful Bari and Yazeed Alnumay and Norah A. Alzahrani and Nouf M. Alotaibi and Hisham Abdullah Alyahya and Sultan AlRashed and Faisal Abdulrahman Mirza and Shaykhah Z. Alsubaie and Hassan A. Alahmed and Ghadah Alabdul-jabbar and Raghad Alkhathran and Yousef Almushayqih and Raneem Alnajim and Salman Alsubaihi and Maryam Al Mansour and Saad Amin Hassan and Majed Alrubaian and Ali Alammari and Zaki Alawami and Abdulmohsen Al-Thubaity and Ahmed Abdelali and Jeril Kuriakose and Abdalghani Abujabal and Nora Al-Twairesh and Areeb Alowisheq and Haidar Khan. 2025. *ALLaM: Large Language Models for Arabic and English*. Model family used in experiments.

DeepSeek-AI. 2024. *DeepSeek LLMs: Scaling Open-Source Language Models with Corpus and Methodology Innovations*. Model family used in comparisons.

AbdelRahim Elmadany and El Moatez Billah Nagoudi and Muhammad Abdul-Mageed. 2023. *Octopus: A Multitask Model and Toolkit for Arabic Natural Language Generation*. Association for Computational Linguistics. Toolkit used.

Inception. 2024. *Jais Family Model Card*. Model card.

Zhiqing Jiang and others. 2023. *Mistral: Fast and Efficient Language Models*. Model family used in comparisons.

Meta AI. 2024. *Llama 3 Model Card*. Model card.

OpenAI. 2024. *GPT-4o System Card*. Model/System card used in experiments.

Ahmed Mahdi Salem. 2024. *Qutoof Min Al-Amthal Al-Yemeniya*. Arab Democratic Center for Strategic, Political, and Economic Studies.

Neha Sengupta and Sunil Kumar Sahu and Bokang Jia and Satheesh Katipomu and Haonan Li and Fajri Koto and William Marshall and Gurpreet Gosal and Cynthia Liu and Zhiming Chen and Osama Mohammed Afzal and Samta Kamboj and Onkar Pandit and Rahul Pal and Lalit Pradhan and Zain Muhammad Mujahid and Massa Baali and Xudong Han and Sondos Mahmoud Bsharat and Alham Fikri Aji and Zhiqiang Shen and Zhengzhong Liu and Natalia Vassilieva and Joel Hestness and Andy Hock and Andrew Feldman and Jonathan Lee and Andrew Jackson and Hector Xuguang Ren and Preslav Nakov and Timothy Baldwin and Eric Xing. 2023. *Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models*. Model used in comparisons.