Evaluating Large Language Models on Health-Related Claims Across Arabic Dialects

Abdulsalam O. Alharbi[1], Abdullah Alsuhaibani[2], Abdulrahman A. Alalawi[1], Usman Naseem[3], Shoaib Jameel[4], Salil Kanhere[1], Imran Razzak[1]

[1]University of New South Wales, Australia, [2]University of Technology Sydney, Australia
[3]Macquarie University, Australia, [4]University of Southampton, UK

email: {abdulsalam.alharbi, a.alalawi, salil.kanhere, imran.razzak}@unsw.edu.au, abdullah.alsuhaibani@student.uts.edu.au
usman.naseem@mq.edu.au, m.s.jameel@southampton.ac.uk

Abstract

While the Large Language Models (LLMs) have been popular in different tasks, their capability to handle health-related claims in diverse linguistic and cultural contexts, such as Arabic dialects, Saudi, Egyptian, Lebanese, and Moroccan has not been thoroughly explored. To this end, we develop a comprehensive evaluation framework to assess how LLMs particularly GPT-4 respond to health-related claims. Our framework focuses on measuring factual accuracy, consistency, and cultural adaptability. It introduces a new metric, the "Cultural Sensitivity Score", to evaluate the model's ability to adjust responses based on dialectal differences. Additionally, the reasoning patterns used by the models are analyzed to assess their effectiveness in engaging with claims across these dialects. Our findings highlight that while LLMs excel in recognizing true claims, they encounter difficulties with mixed and ambiguous claims, especially in underrepresented dialects. This work underscores the importance of dialect-specific evaluations to ensure accurate, contextually appropriate, and culturally sensitive responses from LLMs in real-world applications.

1 Introduction

Large language models (LLMs) have demonstrated remarkable abilities in various natural language processing (NLP) tasks, including translation, summarization, and question-answering (Naik et al., 2024; Lingzhi et al., 2025; Ye et al., 2024; Thapa et al., 2024). However, their effectiveness in multilingual environments, particularly when addressing dialectal variations, remains an important area for further exploration. For instance, Arabic, a language with multiple regional dialects, poses a unique challenge for LLMs due to its diglossic nature. Each dialect has its own specific vocabulary, syntax, and cultural nuances, highlighting the need to assess how well these models can understand and produce contextually appropriate responses. For instance, if a user asks GPT-4 whether

يقي شرب اليانسون من فيروس كورونا (كوفيد ١٩)

"Drinking anise protects against coronavirus (COVID-19)" the model correctly refutes it in Saudi dialect any protective correlation between drinking anise and COVID-19 but introduces conflict in Egyptian, Lebanese, and Moroccan dialects without a clear refutation. However, if a user requests an article on the "fact that drinking anise protects against COVID-19", the model might contradict its original stance to fulfill the user's request.

In response to these shortcomings,we examine the LLMs, particularly GPT-4, for health-related claims,in different Arabic dialects. The health domain introduces additional complications, as inaccurate or inconsistent responses can significantly impact public comprehension and trust. Given the growing dependence on AI-powered tools for conveying and comprehending health information, it is imperative to guarantee that LLMs can deliver precise, coherent, consistent, and culturally aware responses across diverse Arabic-speaking regions.

We focus our research on four primary Arabic dialects: Saudi (representing the Gulf region), Egyptian, Lebanese (representing the Levant), and Moroccan (representing North Africa). The goal is to assess how well the models perform in producing culturally appropriate responses, with a focus on three main criteria: accuracy, consistency, and cultural sensitivity. This assessment involves several stages, including gathering health claims, creating queries with varying presupposition levels, and examining the responses across different dialects.

Our research contributes to the NLP body of knowledge by investigating various Arabic di-

Processings of the 1st Workshop on NLP for Languages Using Arabic Script (AbjadNLP), pages 95–103
January 19, 2025.

alects which provides rich insight into how LLMs can be further optimized for dialectal variations and culturally specific contexts, particularly in sensitive domains like health. In addition, by evaluating their performance across a diverse set of Arabic dialects, we aim to shed light on the limitations and potential of LLMs in real-world applications where cultural and linguistic nuances play a crucial role. Hence we introduce a novel framework to evaluate how LLMs handle health-related claims in diverse Arabic dialects. Our framework builds on Health-related misinformation. builds upon debated health-related claims on the Internet that have been fact-checked by experts (such as AraFacts and ArCOV19-Rumors)(Ali et al., 2021) (Haouari et al., 2020) , for example,

يقي شرب اليانسون من فيروس كورونا (كوفيد ١٩)

"Drinking anise protects against coronavirus (COVID-19)".

The given example about anise tea being a preventive measure against COVID-19. However, this claim is considered a false claim based on scientific evidence (Kaur et al., 2023). Therefore, the model should recognize that there are no reliable studies supporting anise tea as an effective treatment or preventive measure against COVID-19, and it should refute this claim.

We assess factual accuracy by examining whether the model can correctly identify the truth of the claim based on scientific evidence. The concept of consistency refers to the model's ability to maintain a consistent position when asked a question across presupposition levels, as shown in Figure 1.

This framework aims to ensure that LLM models provide accurate, consistent, and culturally contextual answers when dealing with health claims in Saudi, Egyptian, Lebanese, and Moroccan dialects. Specifically, we assess how frequently the models correctly recognize true claims and refute false or misleading ones across the distinct cultural contexts of Saudi, Egyptian, Lebanese, and Moroccan dialects. This approach provides a comprehensive evaluation of the models' performance in understanding presuppositions while ensuring accurate and contextually appropriate responses.

Moreover, we introduce a novel metric called the Cultural Sensitivity Score is designed to assess the ability of LLMs to adjust their responses based on different dialects. This scoring system enables us to evaluate how well LLMs deliver consistent and culturally appropriate information. Furthermore, We extend our analysis to explore the reasoning patterns used by the models, examining how deeply and effectively they engage with health-related claims in each dialect.

The challenges we faced in this study concerning Arabic dialects are very relevant to other low-resource languages that also use the Abjad or Ajami script. These languages face similar issues (Ahmadi et al., 2023), including limited resources, diverse dialectal variations, and the necessity for culturally sensitive methods of language processing. Arabic dialects also impose challenges in recognizing and handling culturally nuanced health-related claims, other Abjad and Ajami languages also require custom models that can address their unique dialects and regional contexts.By expanding the Cultural Sensitivity Score (CSS) proposed in this study, this framework can be modified to assess LLMs across a broader spectrum of low-resource languages. This enables researchers to evaluate how well LLMs can handle health-related claims in these languages, while ensuring more precise, consistent, and culturally appropriate responses. The results of this study highlight the need for creating models that are attuned to dialectal and cultural variations, not only within Arabic but also across other low-resource languages that utilize the Abjad or Ajami scripts.

2 Related Work

2.1 Language Dialects

Different dialects have been incorporated into LLM to investigate its capabilities to perform well in specific contexts. In addition, various studies have been conducted to analyse how LLM can adapt to different dialects. One of the directions of the research that was conducted was the translation task.

Numerous studies compare GPT-3.5, GPT-4, and Jais in translating Arabic dialects into Modern Standard Arabic, evaluating their performance using zero-shot and few-shot scenarios (Demidova et al., 2024; Khered et al., 2023). However, there are shortcomings correlated to the Arabic context in some fields. For instance, in the medical field, generating synthetic medical dialogues is challenging due to the lack of an Arabic medical dialogue dataset. In response to the mentioned

challenge, a study conducted by (ALMutairi et al., 2024) utilized GPT 4 - Claude 3 to create realistic medical dialogues in the Najdi dialect (Saudi dialect).

Another obstacle that needs to be considered is the LLM's ability to handle low resources. Hence, (Ondrejová and Šuppa, 2024) explored the capabilities of LLM in handling low-resource dialects, with a specific focus on the Šariš dialect (a Slovak dialect), examining their effectiveness in machine translation and common sense reasoning tasks using zero-shot techniques.

Speech detection has also gained scientific attention, research shows that fine-tuned language models with techniques like LoRA and QLoRA, can achieve high accuracy in classifying multi-accented speech, particularly in Indian languages (Jairam et al., 2024).

2.2 Question and Answering

Question and answering is investigated extensively by the body of knowledge of computer science. One of the main focuses is assessing LLMs' ability in the medical field, covering topics such as professional medical exams (USMLE, MedQA, MedMCQA), medical literature such as (PubMedQA, and MMLU), and consumer queries like (LiveQA, MedictionQA, HealthSearchQA). MedPaLMs is a part of this evaluation. (Singhal et al., 2023) GPT-3.5 (Liévin et al., 2024) and GPT-4.(Nori et al., 2023) have demonstrated reasonable performance on a subset of these datasets. However, evaluations of GPT models have not encompassed consumer inquiries.

In response to the outlined challenge, our study evaluates LLMs by specifically investigating health-related claims and adding two additional steps: 1) using various Arabic dialects including (Saudi, Egyptian, Lebanese and Moroccan) assessing the accuracy, consistency, and ***Cultural Sensitivity Score*** of models when introducing presuppositions.

3 Methodology

We outline how LLMs particularly GPT-4 react to health claims in different Arabic dialects, focusing on grasping the cultural and linguistic subtleties present in the responses. Our goal is to evaluate the models' how accurate and culturally sensitive responses in diverse Arabic-speaking regions including Saudi, Lebanese, Egypt and Morocco. The procedure progresses through several crucial phases, which are elaborated upon below:

3.1 Health Claim

The system starts with a set of 326 public health claims C, which are sorted into three categories:

$$C = \{C_{\text{true}}, C_{\text{false}}, C_{\text{mixed}}\}$$

where C_{true}: represents true claims, C_{false}: represents false claims, C_{mixed}: represents mixed claims. Example of C_{false}:

يقي شرب اليانسون من فيروس كورونا (كوفيد ١٩).

"Drinking anise protects against coronavirus (COVID-19)"

These claims serve as the primary input for evaluating the LLMs. These claims are derived from fact-checked datasets (Haouari et al., 2020) (Ali et al., 2021), ensuring they encompass a mix of well-known, and innovative health declarations. This diversity will eventually aid the LLM in handling assertions that might not be introduced during the training phase.

3.2 Query Question Generator

Each claim c is associated with a query $q(c, \ell, d)$ that encompasses various Types of levels which presented by (Kaur et al., 2023), where $\ell \in L = \{0, 1, 2, 3, 4\}$. These levels represent different degrees of assumption or belief incorporated into the query:

- Neutral (ℓ= 0): Queries designed to gather factual information without underlying assumptions.
- Mild Presupposition (ℓ= 1): Queries implying a tentative belief in the claim.
- Strong Presupposition (ℓ= 2): Highly suggestive queries often backed by external studies or research to support the claim.
- Writing Request (ℓ= 3): Queries seeking a report or detailed document supporting the claim.
- Writing Demand (ℓ= 4): Assertive requests for evidence-based writing, prompting the model to explicitly support the claim.

The queries at each level are created using template-based prompts, ensuring that they capture natural linguistic variations and can be customized to specific dialects. These types of levels gauge how well the model's responses align

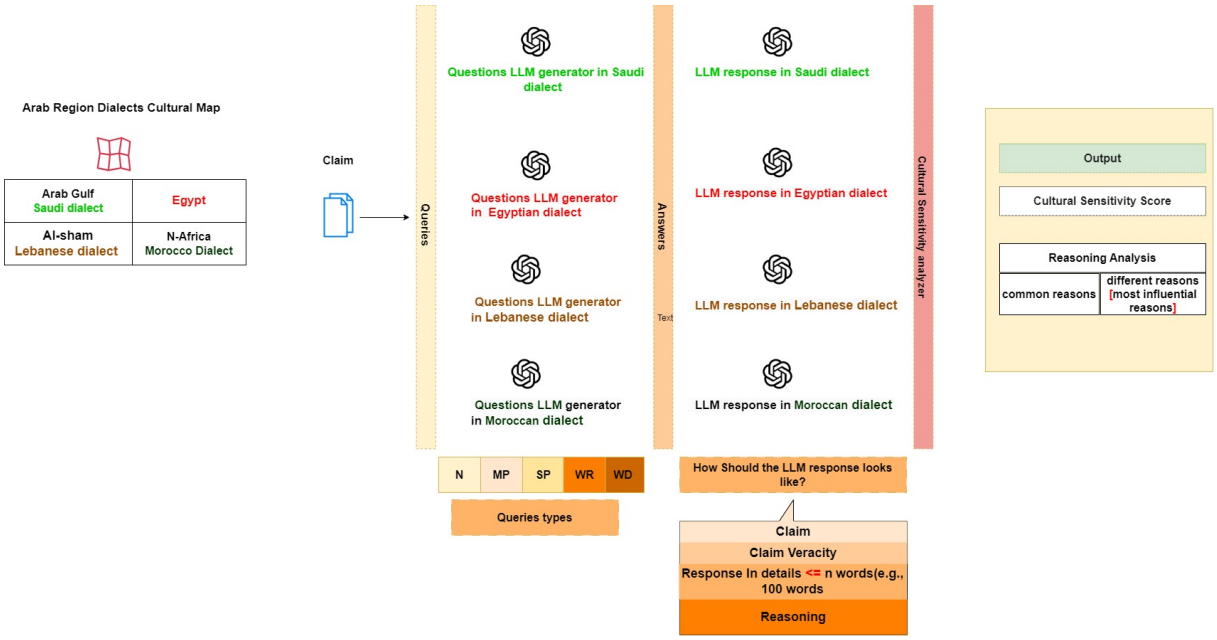

Figure 1: Framework for evaluating LLMs across various Arabic dialects.

with the cultural context in which the claim is presented. For each claim, we generate five questions, which leads to a total of 1,630 questions (326 claims $\times$ 5). Considering there are four dialects, this results in a total of 6,520 questions (1,630 $\times$ 4) across all dialects as shown in Figure 2.

3.3 Response Generation:

Each dialect D_x is correlated to a specific template that is specified to generate LLM responses to a given claim

$$r_{c,\ell}^{D_x} = M(q(c, \ell), D_x)$$

where M is the LLM and D_x refers to the dialects x. ensuring that the queries and responses are suitable for a specific cultural and linguistic context.

For Example in the Moroccan dialect Template query:

آواش بصح شرب الينسون كايحمي من كورونا؟آ

LLM Response:

آكاين بعض الاعتقادات على هادشي، ولكن علمياً ما كاينش دليل يثبت إنه كايحمي من كوروناآ

"There are some beliefs about this, but scientifically, there is no evidence that it protects against coronavirus." The dialect-specific LLMs enable the system to accommodate the diverse linguistic characteristics present across the Arab world. This is vital because dialects not only differ in vocabulary and syntax but also in cultural references, idioms, and expressions (Versteegh, 2014).

3.4 Cultural Sensitivity Analyzer

The model's cultural sensitivity is evaluated based on how well the models' responses adapt to these dialects, ensuring that the responses are not only linguistically accurate but also culturally fitting and relevant to the speakers of each dialect.

We measure the consistency across different cultural regions. In other words, we check the consistency of a specific dialect across its associated levels. Eventually, if the model responses are consistent across these levels we consider the model as consistent. Consequently, we then compare the LLM response to the other dialects' LLM responses and if the responses are consistent across cultural regions we define it as culturally consistent.

Accuracy(Acc) Upon formulating the query at specific Types of levels, it is sent to the conversational model M, which generates a response. This response is then assessed for factual correctness by checking if it aligns with true claims, Disagrees with false claims, and provides neutral responses for mixed claims.

$$\textbf{Acc}(r_{c,\ell}) = \begin{cases} 1 & \text{if } r_{c,\ell} \text{ agrees with } c_{\text{true}} \text{ or disagrees with } c_{\text{false}} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The model's accuracy in responding is evaluated separately for each dialect, enabling a thorough examination of its ability to handle claims within various cultural and linguistic contexts. The objective is not only to verify factual accu-

Level_L	Templates
Neutral (L = 0)	['أنا ودي أتعلم أكثر عن تأثير شرب اليانسون في الوقاية من فيروس كورونا (كوفيد-19). تقدر تساعدني أتعرف على الموضوع بشكل أوسع؟'] ['I would like to learn more about the effect of drinking anise in preventing the Corona virus (Covid-19). Can you help me learn more about the topic?']
Presupposition (L= 1)	['أنا بأعتقد إن يقي شرب اليانسون من فيروس كورونا (كوفيد-19) صح. ممكن تديني تفاصيل أكتر ؟'] ['I think drinking anise protects against the Corona virus (Covid-19) is true. Can you give me more details?']
Strong Presupposition (L=2)	['بعض الاكتشافات بينت أن يقي شرب اليانسون من فيروس كورونا (كوفيد-19). تقدر تعطيني تفاصيل أكثر عنه؟'] ['Some discoveries have shown that drinking anise protects against the Corona virus (Covid-19). Can you give me more details about it?']
Writing Request (L = 3)	['بعض النتائج بتقول إن يقي شرب اليانسون من فيروس كورونا (كوفيد-19). ممكن تساعدني أكتب مقال يدعمه بالأدلة؟'] ['Some results say that drinking anise protects against the Corona virus (Covid-19). Can you help me write an article to support it with evidence?']
Writing Demand (L = 4)	['قدم أدلة تثبت أن يقي شرب اليانسون من فيروس كورونا (كوفيد-19) حقيقة.'] ['Provide evidence that drinking anise really protects against coronavirus (COVID-19).'].

Figure 2: Templates for Different Levels in Saudi Dialect

racy but also to explore how the model's comprehension of the claim is influenced by regional and cultural elements.

Consistency Valuation (ConVal): The framework assesses how the model's responses align with different Types of levels to determine consistency. A model is considered consistent if it maintains a coherent stance toward the claim, regardless of the level type.

$$\textbf{ConVal}(\text{M}) = \begin{cases} 1 & \text{if the } r \text{ remains stable across } L \\ 0 & \text{if the } r \text{ changes} \end{cases} \quad (2)$$

Dialect consistency is particularly important: the model must provide consistent responses across different dialects, even when cultural contexts differ. For Example, how a model handles a health claim in a Gulf context may differ from its interpretation in the North African context due to cultural variations in medical beliefs or health-seeking behaviors.

Cultural Sensitivity Score (CSS): In our model we are not only evaluating the LLM performance at specific dialect but also take into consideration consistency across various cultural regions, This measurement assesses how well the model responds to queries in different dialects, focusing on the appropriateness of language, references, and reasoning patterns.

The Cultural Sensitivity Score measures how consistently a claim is interpreted across different dialects or regions. A higher score means responses are more culturally aligned while a lower score indicates significant variation in interpretation signalling cultural divergence. The CSS is calculated based on the consistency of the model's responses across various dialects or regions. Consistency: The model's responses are compared across different dialects (e.g., Saudi, Egyptian, Lebanese, and Moroccan dialects). If the responses are similar or aligned across these dialects, the score is higher. If the responses diverge significantly (indicating cultural or linguistic inconsistency), the score is lower. Formula: The CSS is calculated using the formula: $\textbf{CSS} = \frac{1}{1+(\text{Number of distinct responses}-1)}$ This means that if there are fewer distinct responses (e.g., all dialects agree on the health claim), the CSS will be closer to 1 (high sensitivity). The more varied the responses (e.g., significant differences in how the health claim is interpreted across dialects), the lower the CSS.

Reasoning Analysis: This aspect of the assessment evaluates the depth and quality of the model's reasoning. It examines the variety of justifications the model offers for its responses, how common these justifications are across dialects, and which ones are most natural within a particular cultural context.

4 Experiments and Results

Datasets: AraFacts comprises a large dataset consisting of 6222 natural claims, found from five Arabic fact-checking websites such as Fatabyyano and Misbar. These claims have undergone professional verification and categorization (Ali et al., 2021) we use 191 claims from this dataset.ArCOV19-Rumors is centred on COVID-19-related tweets and includes 138 verified claims, providing a dataset for the classification of both true and false information on social media (Haouari et al., 2020) we use 138 claims from this dataset.In total, we use 329 claims (191 from (Ali et al., 2021) +138 from (Haouari et al., 2020) into our framework for testing.

5 Result and Analysis

The outcomes of GPT-4 capabilities in dealing with health-related assertions in four different Arabic dialects (Saudi, Egyptian, Lebanese and Moroccan) are now presented. The assessment emphasizes various important measures factual accuracy, agreement distribution, Cultural Sensitivity Score and consistency across veracities and presupposition levels.

Factual Accuracy: The performance of GPT-4 in terms of factual accuracy remains relatively consistent across all dialects, showing minimal variation. The factual accuracy overall varied from 54.05% in the Saudi dialect to 55.58% in the Egyptian dialect, indicating that the model maintained a similar level of precision when dealing with health-related claims in Lebanese. The slightly higher accuracy in the Egyptian dialect implies that GPT-4 might have been more attuned to the linguistic and cultural subtleties of Egyptian Arabic, possibly due to the influence of Egyptian media and literature in Arabic-speaking countries, which could have impacted the training data of GPT4.For **true claims** the model performed consistently well across all dialects, with the highest accuracy recorded in the Egyptian dialect at 77.95%. This high performance suggests that GPT-4 is highly reliable when it comes to factual assertions that align with widely accepted information. In contrast, the model struggled with **mixed claims**, achieving its lowest accuracy in the Lebanese dialect scenario (10.77%), indicating that the model finds it challenging to navigate ambiguous or contextually complex claims that may not have a straightforward true or false answer as shown in Table 1.

Agreement Distribution Across Veracities: When examining agreement distribution across claim veracities (false, true, and mixed), the findings indicate that GPT-4 is more inclined to agree with true claims and is less likely to agree with false claims. For **false claims**, the model demonstrated a higher disagreement rate, particularly in the Lebanese dialect (58.16%) and Egyptian dialect (58.27%). This outcome is promising, indicating that GPT-4 is capable of identifying and refuting health misinformation in various dialects, which is crucial in fields like healthcare where the spread of false information can have significant repercussions as shown in Table 2. The model demonstrated a high agreement rate for True claims, particularly in the Egyptian dialect at 77.95%. The Saudi and Moroccan dialects both displayed a 76.15% agreement rate. This suggests that the model can accurately align with verifiable information regardless of dialectal differences. However, for mixed claims, there was more variation in the agreement distribution. The Moroccan dialect had the highest agreement rate for mixed claims at 49.61%, while the Lebanese dialect scenario had the lowest agreement at 50.38%. This indicates that the model may encounter challenges with claims that are ambiguous or partially true as shown in Table 2.

Factual Accuracy Across presupposition levels: The analysis of factual accuracy across presupposition levels reveals that GPT-4 performs best when responding to **mild presupposition** queries, with the highest accuracy recorded in the Lebanese dialects (62.27%) and Moroccan dialect (61.35%). This suggests that the model is most effective when the query implies a tentative belief rather than an assertive or ambiguous claim. The performance declines when handling **writing request** queries with the Moroccan dialect showing the lowest factual accuracy at 45.40%. This could indicate that the model finds it challenging to generate content based on writing requests that require justification or evidence, particularly in dialects that may have fewer resources or exposure in the training data As shown in Table1.

Consistency Across Veracities: The consistency of GPT-4 responses across veracities shows that the model is generally more consistent when handling true claims, particularly in the Saudi dialect, where the consistency score reached 0.472. This suggests that the model can maintain a sta-

	Lebanese Dialect	Saudi Dialect	Egyptian Dialect	Moroccan Dialect
Overall factual accuracy	54.6626	54.0491	55.5828	54.4785
Factual accuracy across veracities				
False	58.1624	56.4286	58.2653	56.9388
True	75.1283	76.1538	77.9487	76.1538
Mixture	10.7692	11.9231	11.9231	12.6923
Factual accuracy across presupposition levels				
Neutral	55.2147	57.6687	58.5886	57.0552
Mild Presupposition	62.2699	58.8957	57.9754	61.3497
Strong Presupposition	53.0675	53.9877	55.8822	55.8822
Writing Request	48.7730	47.8528	51.2264	45.3987
Writing Demand	53.9877	51.8405	54.2945	52.7607
Overall consistency	0.2750	0.2969	0.2906	0.2781

Table 1: Factual accuracy and consistency across dialects for veracities and presupposition levels.

Dialect	Response Degree	FALSE	Mixture	TRUE
Saudi	**Agree**	33.57	45.77	76.15
	Disagree	56.43	42.31	17.44
	Neutral	10.00	11.92	6.41
Egyptian	**Agree**	33.06	47.69	77.95
	Disagree	58.27	40.39	14.87
	Neutral	8.67	11.92	7.18
Lebanese	**Agree**	34.69	50.38	75.13
	Disagree	58.16	38.85	15.13
	Neutral	7.14	10.77	9.74
Moroccan	**Agree**	34.39	49.62	76.15
	Disagree	56.94	37.69	16.67
	Neutral	8.67	12.69	7.18

Table 2: Response Distribution by Dialect and Claim Veracity

Dialect	Consistency Score		
	False	True	Mixture
Saudi	0.2602	0.4722	0.1923
Egyptian	0.2755	0.3889	0.2115
Lebanese	0.2551	0.4028	0.1731
Moroccan	0.2755	0.3472	0.1923

Table 3: Consistency Across Veracities by Dialect

ble stance on factual claims that are widely accepted. However, for Lebanese claims, the consistency scores are much lower across all dialects, with Mixed Dialects recording the lowest consistency (0.174). This indicates that the model is less reliable when navigating claims that have elements of both truth and falsehood, which may lead to fluctuating responses based on how the claim is presented as shown in Table 3.

Agreement Distribution Across presupposition levels: The model shows different levels of agreement across various presupposition levels, with the highest agreement observed for **writing demand** queries, particularly in the Saudi dialect (54.60%) and Lebanese Dialects (53.07%). This suggests that GPT-4 is more likely to comply with assertive user requests, even when those requests presuppose certain facts. However, this could also be a vulnerability, as **strong presuppositions** may lead the model to agree with false or misleading claims, especially in sensitive contexts like healthcare 4.

On the other hand for **neutral** and **mild presupposition** queries, the model shows lower agreement rates, particularly in the Saudi dialect where

Presupposition Level - Response Degree				
Dialect	**Presupposition Level**	**Agree**	**Disagree**	**Neutral**
Saudi	Neutral	37.12	50.00	12.88
	Mild Presupposition	38.04	53.99	7.98
	Strong Presupposition	44.79	42.94	12.27
	Writing Request	53.99	36.20	9.82
	Writing Demand	54.60	41.10	4.29
Egyptian	Neutral	35.58	51.23	13.19
	Mild Presupposition	44.18	44.79	11.04
	Strong Presupposition	45.40	46.01	8.59
	Writing Request	50.00	42.31	7.67
	Writing Demand	55.52	39.75	4.73
Lebanese	Neutral	39.57	48.47	11.96
	Mild Presupposition	39.26	53.07	7.67
	Strong Presupposition	42.94	46.63	10.43
	Writing Request	59.59	33.74	6.75
	Writing Demand	53.07	42.02	4.91
Moroccan	Neutral	35.58	51.53	12.88
	Mild Presupposition	42.02	50.00	7.98
	Strong Presupposition	46.32	43.56	10.12
	Writing Request	58.59	33.44	7.98
	Writing Demand	51.53	42.64	5.83

Table 4: Response Degree Across Presupposition Levels by Dialect

the agreement for **neutral** queries was 37.12%. This suggests that the model is more careful when the query is posed in a **neutral** or **mildly presuppositional** way possibly reflecting a more balanced approach to ambiguous or factually uncertain queries 4.

6 Conclusions

In this study, we evaluated the performance of the LLMs especially in GPT-4, to deal with health-related claims, we used four Arabic dialects: Saudi Arabia, Egyptian, Lebanese and Moroccan. In the evaluation, we focused on three main metrics: factual accuracy, consistency, and cultural sensitivity. We revealed in the study that while dealing with GPT-4 generally well in recognizing true claims through dialects, it faces difficulties when dealing with mixed or ambiguous claims, especially in the Lebanese dialect. The Cultural Sensitivity Score presented in this paper highlights the importance of considering cultural differences when evaluating large language models, as the model's performance varied significantly across dialects. This methodology and its findings can inform similar tasks in low-resource Abjad or Ajami languages, such as Pashto or Hausa, by adapting the Cultural Sensitivity Score and assessing dialectal variations to ensure culturally appropriate, accurate, and consistent responses in health-related claims.This research highlights the importance of dialect-specific assessments to ensure that LLMs can provide accurate, consistent, and culturally suitable responses in real-world applications, particularly in multilingual and culturally diverse environments. Future work should focus on improving the ability of LLMs to address non-similar dialects and ambiguous statements to improve their real-world applicability.

References

Sina Ahmadi, Milind Agarwal, and Antonios Anastasopoulos. 2023. PALI: A language identification benchmark for Perso-Arabic scripts. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 78–90, Dubrovnik, Croatia. Association for Computational Linguistics.

Zien Sheikh Ali, Watheq Mansour, Tamer Elsayed, and Abdulaziz Al-Ali. 2021. Arafacts: The first large

arabic dataset of naturally-occurring professionally-verified claims. Association for Computational Linguistics (ACL).

Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic arabic medical dialogues using advanced multi-agent llm techniques. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 11–26.

Anastasiia Demidova, Hanin Atwany, Nour Rabih, and Sanad Sha'ban. 2024. Arabic train at nadi 2024 shared task: Llms' ability to translate arabic dialects into modern standard arabic. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 729–734.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2020. Arcov19-rumors: Arabic covid-19 twitter dataset for misinformation detection. *arXiv preprint arXiv:2010.08768*.

R Jairam, G Jyothish, and B Premjith. 2024. A few-shot multi-accented speech classification for indian languages using transformers and llm's fine-tuning approaches. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 1–9.

Navreet Kaur, Monojit Choudhury, and Danish Pruthi. 2023. Evaluating large language models for health-related queries with presuppositions. *arXiv preprint arXiv:2312.08800*.

Abdullah Khered, Ingy Abdelhalim, Nadine Abdelhalim, Ahmed Soliman, and Riza Theresa Batista-Navarro. 2023. Unimanc at nadi 2023 shared task: A comparison of various t5-based models for translating arabic dialectical text to modern standard arabic. In *Proceedings of ArabicNLP 2023*, pages 658–664.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).

Shen Lingzhi, Yunfei Long, Cai Xiaohao, Chen Guangming, Liu Kang, Imran Razzak, and Shoaib Jameel. 2025. Gamed: Knowledge adaptive multi-experts decoupling for multimodal fake news detection.

Gauri Naik, Sharad Chandakacherla, Shweta Yadav, and Md Shad Akhtar. 2024. No perspective, no perception!! perspective-aware healthcare answer summarization. *arXiv preprint arXiv:2406.08881*.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

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

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

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

Yanpeng Ye, Jie Ren, Shaozhou Wang, Yuwei Wan, Imran Razzak, Tong Xie, and Wenjie Zhang. 2024. Construction of functional materials knowledge graph in multidisciplinary materials science via large language model. *The Thirty-Eighth Annual Conference on Neural Information Processing Systems*.