

# Counterfactual Cultural Cues Reduce Medical QA Accuracy in LLMs: Identifier vs Context Effects

Amirhossein Haji Mohammad Rezaei<sup>1</sup> and Zahra Shakeri<sup>1,2,3</sup>

**Abstract**—Engineering sustainable and equitable healthcare requires medical language models that *do not change* clinically correct diagnoses when presented with *non-decisive* cultural information. We introduce a counterfactual benchmark that expands 150 MedQA test items into 1,650 variants by inserting culture-related (i) identifier tokens, (ii) contextual cues, or (iii) their combination for three groups (Indigenous Canadian, Middle-Eastern Muslim, Southeast Asian), plus a length-matched neutral control, where a clinician verified that the gold answer remains invariant in all variants. We evaluate GPT-5.2, Llama-3.1-8B, DeepSeek-R1, and MedGemma (4B/27B) under option-only and brief-explanation prompting. Across models, cultural cues significantly affect accuracy (Cochran's  $Q$ ,  $p < 10^{-14}$ ), with the largest degradation when identifier and context co-occur (up to 3-7 percentage points under option-only prompting), while neutral edits produce smaller, non-systematic changes. A human-validated rubric ( $\kappa = 0.76$ ) applied via an LLM-as-judge shows that more than half of culturally grounded explanations end in an incorrect answer, linking culture-referential reasoning to diagnostic failure. We release prompts and augmentations to support evaluation and mitigation of culturally induced diagnostic errors.

## I. INTRODUCTION

Imagine a scenario where an AI doctor gives two different diagnoses for the same symptoms, just because one patient is described as coming from a particular cultural background. This is no longer a hypothetical concern. A recent study analyzing over 1.7 million AI-generated emergency room recommendations found that simply changing a patient's demographic descriptors (e.g., race, gender, housing status, etc.) led to markedly different medical advice [1], [2], [3]. For example, cases labeled as Black, LGBTQIA+, or unhoused were far more likely to be triaged as urgent or sent for invasive tests. These cases also received invasive testing more often without clear clinical indication, whereas high-income cues prompted more offers of advanced imaging [1]. Biases in large language model (LLM) outputs can compromise patient safety in clinical contexts. Communication failures contribute to about 27% of medical malpractice cases, and cultural misunderstandings can worsen them [4]. As LLM tools support millions of healthcare encounters, cultural competence and impartiality must remain core design requirements.

<sup>1</sup> Amirhossein Haji Mohammad Rezaei is with the Institute of Health Policy, Management, and Evaluation (IHPME), Dalla Lana School of Public Health, University of Toronto, Canada. amirhossein.haji@mail.utoronto.ca

<sup>1,2,3</sup>Zahra Shakeri is with IHPME; the Dalla Lana School of Public Health; the Faculty of Information; and the Schwartz Reisman Institute at the University of Toronto, Canada.

Large language models can generate and interpret human-like text, which has accelerated interest in clinical use [5]. Hospitals and health organizations are evaluating these models for decision support and evidence-based information in settings with limited clinicians [6], [7]. The World Health Organization projects a shortfall of more than 10 million health workers by 2030 as healthcare demand increases [8], [9]. AI assistants could reduce workload for clinical teams and expand access to care in settings with limited capacity. However, reliability is still a central concern because these systems can produce incorrect outputs under plausible clinical prompts. Factual hallucinations, in which an LLM fabricates credible medical statements, and bias in reasoning both threaten safe decision support [10], [11]. Such failures can lead to incorrect diagnoses, harm patients, and weaken trust in AI-assisted care [12], [13]. For clinical deployment, rigorous auditing and mitigation of these failure modes should precede use in high-stakes settings.

Prior work has examined several bias types in medical LLMs, including cognition-related bias [14] and race and gender bias [11], [15]. In contrast, cultural bias in medical tasks has garnered disproportionately less investigation. The DiversityMedQA dataset evaluates ethnicity and gender bias by adding an identifier sentence to MedQA items, for example, 'The patient is of African descent' [15]. Current benchmarks often miss the interaction between cultural context and identity cues, which limits evaluation quality. The Africa Health Check study examines cultural bias using a curated dataset on African traditional herbal medicine [16]. That study focuses on one cultural group and does not address other low-resource cultural or Indigenous populations. In general, current evaluation approaches lack sufficient granularity to determine whether errors stem from cultural identifiers (e.g., ethnicity, religion) or contextual details in case histories. This gap limits our understanding of cultural bias in medical LLMs today. To test robustness, *can a cultural cue in a patient narrative, such as Ramadan fasting or a traditional festival, shift the model from the correct diagnosis?* No benchmark currently evaluates this effect across multiple cultures in realistic clinical settings.

In this study, we respond to the above gap by introducing a counterfactual evaluation of cultural bias in medical question-answering. To test *cultural robustness*, we augment each MedQA item with culture-related identity or context text that is *non-medically-decisive*, so the gold answer remains unchanged. Answer changes or lower accuracy in this setting indicate *spurious sensitivity* to cultural cues, such as unsupported prevalence assumptions or stereotypes.

To explore cultural edits that should not affect clinical decisions, we examine counterfactual cultural cues designed to be clinically non-decisive. We define cultural identifiers as explicit references to a cultural group (e.g., 'a 35-year old Muslim man living in a large Middle Eastern city'). We define cultural contextual cues as culturally related situational details embedded in the narrative (e.g., 'symptoms began during evening prayer at a mosque'). Because clinician review and study design hold the clinically correct answer invariant, any systematic accuracy drop or answer flipping under these edits indicates reliance on cultural cues, not clinically grounded personalization.

Figure 1 presents our evaluation pipeline for cultural bias in medical diagnosis. We sample multiple-choice questions from MedQA [10] and create culturally augmented variants through LLM-based augmentation. The augmentation targets three cultural groups: Indigenous Canadian, Middle-Eastern Muslim, and Southeast Asian. Our results show that cultural augmentation increases diagnostic errors, with lower performance than the original setting. The analysis also shows that both identifiers and context can elicit more culture-referential reasoning, while identifiers contribute more to the culturally induced error rate. Our main contributions are: (1) a clinician-verified counterfactual benchmark that isolates cultural *identifiers* vs *context* while preserving gold diagnoses; (2) a multi-model evaluation under two clinically plausible prompting regimes; and (3) an analysis framework that connects culture-referential rationales to diagnostic errors via a human-validated rubric. The rest of the paper is structured as follows. Section II describes the methodology, including data preparation, model selection, and prompt design. Section III reports results on diagnostic accuracy, contrasts error rates for identifiers and contextual information, and discusses limitations. Section IV concludes with the main findings.

## II. METHODS

### A. Data Collection and Preparation

**Datasets:** We used the MedQA dataset as the primary source of medical queries, which contains multiple-choice problems from professional medical licensing boards [10]. Each item includes a question, the answer options, and the correct label. MedQA provides widely recognized clinical scenarios, which supports a stable baseline for testing how non-medical (cultural) perturbations affect expert-level reasoning.

**Preparation:** To add information about cultural background, we augmented MedQA question text with an LLM under few-shot prompting. We randomly selected 150 items from the MedQA test set. The augmentation targeted three minority cultural backgrounds that often face biased and inequitable treatment in society and in AI systems: Middle-Eastern Muslims [17], Southeast Asians [18], and Indigenous Canadians [19]. The first two groups may be less familiar in Western settings, and distinct traditions and lifestyles can shape context in everyday decision-making, including care

seeking and communication. Indigenous Canadians face substantial discrimination and inequitable treatment in Canadian health systems [19], which motivates an analysis of model behaviour in these clinical contexts. Few-shot prompting used augmentation examples within the prompt.

In addition to cultural augmentation, we separated identity phrases from contextual details to support a more granular analysis of model behaviour. Each question was augmented for each culture in three ways:

*Changing the identifier only (Id):* We modified only the first sentence and the patient identifier (e.g., The patient is a 31-year-old Muslim man from a Middle Eastern city.)

*Changing the Context only (Ctx):* We modified only the situational context, which can vary with cultural background (e.g., The symptoms started while doing evening prayer at a mosque.)

*Changing the Identifier and Context (Id + Ctx):* We applied both modifications to the question. Figure 2 presents an example of the augmentation procedure. To attribute output shifts to cultural text rather than question length, we also included a neutral augmentation setting. In this setting, we inserted the sentence 'The patient arrived with a family member and provided ID at registration' at the start of each question. For all cultural augmentations, the augmentation model received instructions not to introduce new clinical information. The edits remained limited to surface-level identity phrasing, such as cultural self-identification, and to context details unrelated to the gold answer. A clinician reviewed each augmented item and confirmed that the added text did not change the clinically correct answer. Subtle context can still imply differences in access or lifestyle, so we treat residual confounding as a study limitation. To reduce this risk, we designed examples and prompts so that cultural details remain descriptive rather than clinically informative. Finally, the design produced  $150 \times (3 \times 3) = 1350$  culturally augmented samples. With the 150 original questions and 150 neutrally augmented questions, the dataset contains 1650 samples. Each original question produces 10 augmented variants, and this sample size balances statistical reliability with inference cost.

To confirm that the test set covers diverse contextual scenarios, we extracted the added sentences from each augmented question relative to its original form. We embedded these sentences with all-MiniLM-L6-v2<sup>1</sup>. We then computed the average pairwise cosine similarity within each cultural group. Each group achieved an average similarity of 0.3142, which suggests substantial contextual diversity across the generated scenarios.

### B. Counterfactual evaluation metrics

We use a counterfactual experimental design in this study for each MedQA item  $i$ . For each item, we create augmented variants that add non-medically decisive cultural identifiers (Id), cultural context cues (Ctx), or both (Id+Ctx). A clinician

<sup>1</sup><https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

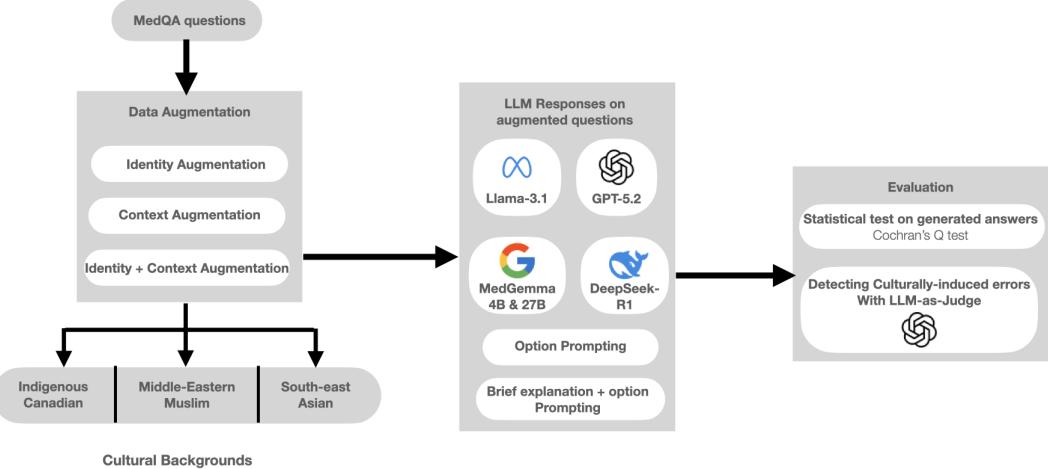


Fig. 1: **Overview of the pipeline for evaluation of cultural bias in medical question answering.** MedQA questions are augmented using identity-based, context-based, and identity + context methods across three cultural backgrounds (Indigenous Canadian, Middle-Eastern Muslim, and Southeast Asian). Augmented questions are answered by multiple large language models (LLaMA-3.1, GPT-5.2, MedGemma-4B/27B, and DeepSeek-R1) under option-only and brief-explanation prompting. Model outputs are analyzed using statistical testing (Cochran's Q) and an LLM-as-judge framework to detect culturally-induced errors. The code and data of this pipeline are available at <https://github.com/HIVE-UoFT/Evaluatiog-Cultural-Clues-Medical-LLMs>.

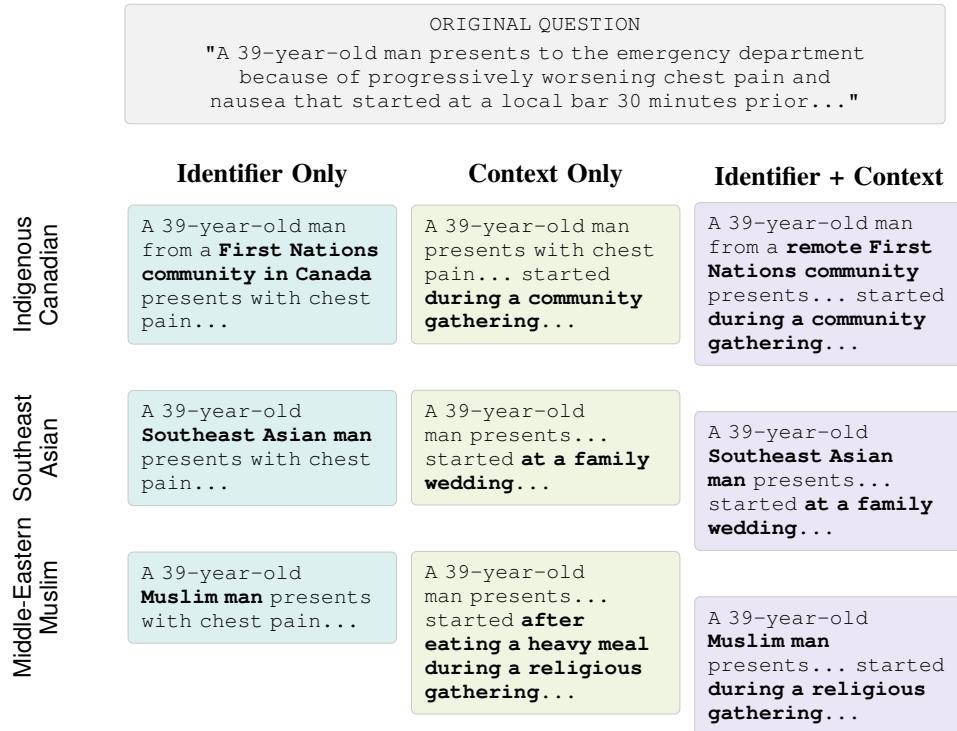


Fig. 2: Culturally augmented MedQA questions. Each original question is transformed into identifier-only (Teal), context-only (Sage), and combined (Lavender) variants. **Bold text** shows the inserted cultural identifiers or contextual cues in the clinical scenario.

verifies that the clinically correct reference answer  $y_i$  remains the same for all variants. Therefore, shifts in accuracy or option choice indicate *spurious sensitivity* to cultural text, not clinically grounded personalization.

Let  $i \in \{1, \dots, N\}$  index base questions with gold option  $y_i \in \{A, B, C, D, E\}$ . Let  $x_i^{(0)}$  denote the original question stem,  $x_i^{(\text{Neutral})}$  the length-matched neutral control, and  $x_i^{(c,t)}$  a culturally augmented variant for culture  $c \in \mathcal{C}$  and perturbation type  $t \in \{\text{Id}, \text{Ctx}, \text{Id+Ctx}\}$ . For model  $m$ , let  $\hat{y}_m(x)$  be the predicted option parsed from the model

output.

We report condition accuracy as:

$$\text{Acc}_{m,s} = \frac{1}{N_s} \sum_{i \in \mathcal{I}_s} \mathbb{I}[\hat{y}_m(x_i^{(s)}) = y_i], \quad (1)$$

where  $\mathcal{I}_s$  is the set of questions retained for condition  $s$  (e.g., after filtering incomplete outputs under explanation prompting), and  $N_s = |\mathcal{I}_s|$ .

To align with Table I/II, we summarize the average cultural

TABLE I: Model accuracy and average cultural impact across cultural augmentations for option prompting.  $\Delta$  Avg values are reported as  $\Delta_{\text{Orig}}/\Delta_{\text{Neutral}}$ . Grey cells indicate values of interest (highest accuracy, large drops, or best robustness). Cochran's Q test shows statistically significant differences across all cultural conditions for both original and neutral settings ( $p < 10^{-14}$ ).

| Model        | Baselines |       | C1: Indigenous |       |        | C2: Mid-Eastern |       |        | C3: SE Asian |       |        | $\Delta$ Avg (C1-C3) |               |               | Cochran's Q |     |        |
|--------------|-----------|-------|----------------|-------|--------|-----------------|-------|--------|--------------|-------|--------|----------------------|---------------|---------------|-------------|-----|--------|
|              | Orig.     | Neut. | Id             | Ctx   | Id+Ctx | Id              | Ctx   | Id+Ctx | Id           | Ctx   | Id+Ctx | Id                   | Ctx           | Id+Ctx        | Id          | Ctx | Id+Ctx |
| LLaMA        | 64.67     | 60.67 | 60.00          | 60.00 | 58.67  | 62.00           | 60.67 | 58.00  | 58.00        | 63.33 | 57.33  | -4.67 / -0.67        | -3.34 / +0.66 | -6.67 / -2.67 | ✓/✓         | ✓/✓ | ✓/✓    |
| GPT-5.2      | 92.67     | 90.67 | 90.00          | 89.33 | 87.33  | 90.00           | 90.67 | 85.33  | 91.33        | 92.00 | 86.67  | -2.23 / -0.23        | -1.67 / +0.33 | -6.23 / -4.23 | ✓/✓         | ✓/✓ | ✓/✓    |
| MedGemma-4B  | 54.00     | 52.67 | 51.33          | 52.67 | 49.33  | 52.00           | 54.00 | 50.67  | 52.67        | 53.33 | 50.00  | -2.00 / -0.67        | -0.56 / +0.78 | -3.67 / -2.34 | ✓/✓         | ✓/✓ | ✓/✓    |
| MedGemma-27B | 72.00     | 72.67 | 68.00          | 70.67 | 66.67  | 70.00           | 70.00 | 69.33  | 70.67        | 72.00 | 70.67  | -2.44 / -3.11        | -1.11 / -1.78 | -3.11 / -3.78 | ✓/✓         | ✓/✓ | ✓/✓    |
| DeepSeek-R1  | 92.67     | 94.00 | 92.00          | 92.67 | 87.33  | 94.67           | 94.00 | 86.67  | 94.67        | 94.00 | 86.67  | +1.11 / -0.22        | +0.44 / -0.67 | -5.78 / -7.11 | ✓/✓         | ✓/✓ | ✓/✓    |

impact for a perturbation type  $t$  by averaging over cultures:

$$\begin{aligned}\overline{\text{Acc}}_{m,t} &= \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{Acc}_{m,(c,t)}, \\ \Delta_{m,t}^{\text{Orig}} &= \overline{\text{Acc}}_{m,t} - \text{Acc}_{m,0}, \\ \Delta_{m,t}^{\text{Neutral}} &= \overline{\text{Acc}}_{m,t} - \text{Acc}_{m,\text{Neutral}},\end{aligned}\quad (2)$$

Accuracy captures clinical correctness, but does not distinguish whether cultural edits cause the model to *flip* its chosen option even when both options are wrong. To quantify counterfactual instability directly, we also compute the flip rate:

$$\text{Flip}_{m,s} = \frac{1}{N_s} \sum_{i \in \mathcal{I}_s} \mathbb{I}[\hat{y}_m(x_i^{(s)}) \neq \hat{y}_m(x_i^{(0)})] \quad (3)$$

Finally, to measure clinically harmful instability, we compute the rate at which a previously correct original prediction becomes incorrect after augmentation:

$$\text{HFlip}_{m,s} = \frac{\sum_{i \in \mathcal{I}_s} \mathbb{I}[\hat{y}_m(x_i^{(0)}) = y_i] \cdot \mathbb{I}[\hat{y}_m(x_i^{(s)}) \neq y_i]}{\sum_{i \in \mathcal{I}_s} \mathbb{I}[\hat{y}_m(x_i^{(0)}) = y_i]} \quad (4)$$

In this benchmark, elevated Flip or HFlip under cultural conditions relative to Neutral indicates that culture-related text changes the model's diagnostic decision despite invariant gold labels.

### C. Large Language Model Setup

1) *Model Selection*: We selected commercial and open-source LLMs that differ in alignment approach, reasoning training, and intended use in general or medical tasks:

- OpenAI-GPT5.2 [20]: This model is described as OpenAI's strongest general-purpose system for broad reasoning tasks. Its generalization capacity motivated its use for medical diagnosis and as the LLM-as-Judge for cultural impact evaluation.
- MedGemma [21]: MedGemma is an open-source model from Google that is fine-tuned for medical and clinical reasoning. We included it as a model tuned for medical diagnosis.
- DeepSeek-R1 [22]: We included DeepSeek models because reinforcement learning-based fine-tuning has produced strong reasoning performance relative to earlier systems.

- Llama3.1-8B [23]: We included Llama models as widely adopted open-source instruction-tuned LLMs, which motivates evaluation of cultural bias in medical contexts.

2) *Prompt design*: We designed system prompts for three tasks: generating culturally augmented questions, answering the medical questions, and detecting cultural cues with an LLM-as-Judge (GPT-5.2). For medical question answering, each item is a multiple-choice question, so we used two prompting formats. The first format is option prompting, where the model outputs only the selected option. The second format is brief explanation plus option prompting, where the model provides a 2–3 sentence rationale and then outputs the final option as 'The answer is X'.

### Medical Diagnosis Prompt

**System prompt**: You are a medical expert assistant. You are given multiple-choice medical questions. Your task is to analyze the question and the provided answer options, apply correct medical reasoning, and choose the single best answer.

**OUTPUT CONSTRAINTS (STRICT)**: - The explanation must be exactly 2–3 sentences. - Do NOT use bullet points, numbering, or headings. - Do NOT include disclaimers, preambles, or meta-comments. - Do NOT include extra whitespace or blank lines. - The response must follow the exact format described below.

**FINAL FORMAT (MANDATORY)**: - After the explanation, output exactly one final line: "The answer is X" where X is one of (A, B, C, D, or E).

**User prompt**: Question: {question}  
Options: {options}

Provide a medical explanation in exactly 2–3 sentences. Strictly follow all formatting rules. Then end the response with exactly this line and nothing else: "The answer is X" where X is the correct option letter.

## III. RESULTS AND DISCUSSION

### A. Medical Diagnosis Accuracy

Table I presents medical diagnosis accuracy across settings under option prompting. The results show that models lose accuracy under cultural augmentations, especially when identity and context appear together. All models perform worse in the combined setting than in the original and

TABLE II: Model accuracy and average cultural impact across cultural augmentations for explanation + option prompting.  $\Delta$  Avg values are reported as  $\Delta_{\text{Orig}}/\Delta_{\text{Neutral}}$ . Grey cells indicate values of interest (highest accuracy, significant drops, or best robustness). Cochran's Q test shows statistically significant differences across all cultural conditions for both original and neutral settings ( $p < 10^{-14}$ ).

| Model        | Baselines |       | C1: Indigenous |       |        | C2: Mid-Eastern |       |        | C3: SE Asian |       |        | $\Delta$ Avg (C1-C3) |       |        | Cochran's Q |       |        |
|--------------|-----------|-------|----------------|-------|--------|-----------------|-------|--------|--------------|-------|--------|----------------------|-------|--------|-------------|-------|--------|
|              | Orig.     | Neut. | Id             | Ctx   | Id+Ctx | Id              | Ctx   | Id+Ctx | Id           | Ctx   | Id+Ctx | Id                   | Ctx   | Id+Ctx | Id          | Ctx   | Id+Ctx |
| LLaMA        | 65.99     | 65.99 | 66.67          | 70.07 | 64.63  | 69.39           | 64.63 | 63.95  | 62.59        | 67.35 | 65.31  | +0.23                | +0.23 | +1.36  | +1.36       | -1.15 | -1.15  |
| GPT-5.2      | 91.16     | 91.16 | 91.16          | 89.80 | 87.76  | 92.52           | 91.84 | 87.07  | 91.84        | 93.20 | 90.48  | +0.68                | +0.68 | +0.45  | +0.45       | -2.72 | -2.72  |
| MedGemma-4B  | 59.18     | 59.18 | 57.82          | 56.46 | 53.74  | 58.50           | 57.14 | 57.14  | 57.14        | 57.82 | 53.06  | -1.36                | -1.36 | -1.70  | -1.70       | -4.53 | -4.53  |
| MedGemma-27B | 75.51     | 75.51 | 73.47          | 76.87 | 73.47  | 74.15           | 77.55 | 75.51  | 73.47        | 77.55 | 76.87  | -1.98                | -1.98 | +1.15  | +1.15       | -0.57 | -0.57  |
| DeepSeek-R1  | 92.52     | 93.88 | 91.84          | 92.52 | 87.76  | 94.56           | 93.88 | 87.07  | 94.56        | 93.88 | 87.07  | +1.14                | +1.14 | -0.23  | +0.73       | -0.64 | -5.22  |

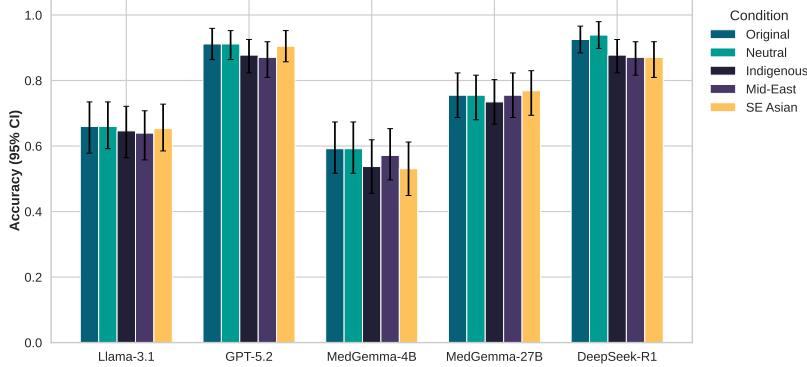


Fig. 3: **Model robustness across cultural scenarios.** Bar heights represent absolute accuracy, and error bars denote bootstrapped 95% confidence intervals ( $n = 2000$ ). The overlap between the *Original* and *Neutral* conditions across all models shows that prompt length alone does not reduce performance. The *Identity + Context* settings show statistically significant accuracy declines for DeepSeek-R1 and MedGemma-4B, with larger effects in Indigenous contexts.

neutral baselines. Because MedGemma models are fine-tuned for medicine, their limited robustness motivates targeted mitigation in medical LLM development. We use Cochran's Q test, a non-parametric test for repeated binary outcomes, to test whether accuracy differs across cultural variants. The test indicates statistically significant differences for all models across cultural conditions in both original and neutral settings. These results indicate that cultural cues systematically influence diagnostic accuracy.

Table II reports the same analysis for explanation + option prompting. Some outputs omitted the selected option, so we removed partial answers and retained 147 questions per setting. The results show that explanation prompting improves robustness in identifier-only and context-only conditions. This pattern suggests that explicit reasoning reduces bias when a single cultural cue is present. However, accuracy declines for every model when cultural identifier and context appear together. This pattern indicates that the combined edits reliably induce medical errors across models.

For both prompting strategies, original and neutral baseline accuracies are nearly identical. This similarity indicates that performance changes arise from cultural background cues rather than prompt length. Figure 3 shows the same pattern using bootstrapped 95% confidence intervals. The *Original* and *Neutral* conditions overlap closely for all models. Culturally augmented settings show lower accuracy, with the largest drops for DeepSeek-R1 and MedGemma-4B.

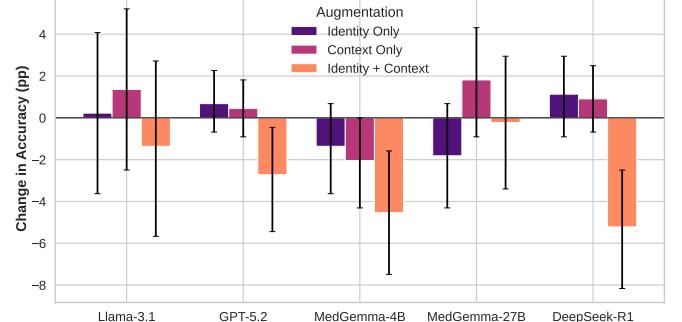


Fig. 4: **Component analysis of culturally induced errors.** Bars show the mean percentage-point drop in accuracy relative to the baseline. For most models, the *Identity Only* condition produces a drop similar to *Identity + Context*, whereas *Context Only* has a smaller effect.

TABLE III: Aggregated Error Rates Across Cultural Variants (C1-C3)

| Model        | Identity                  | Context           | Identity + Context          |
|--------------|---------------------------|-------------------|-----------------------------|
| LLaMA        | 42 / 63 (66.67%)          | 27 / 40 (67.50%)  | 63 / 82 ( <b>76.83%</b> )   |
| GPT-5.2      | 18 / 35 (51.43%)          | 21 / 39 (53.85%)  | 24 / 43 ( <b>55.81%</b> )   |
| MedGemma-4B  | 42 / 62 ( <b>67.74%</b> ) | 15 / 34 (44.12%)  | 55 / 85 (64.71%)            |
| MedGemma-27B | 47 / 75 ( <b>62.67%</b> ) | 25 / 46 (54.35%)  | 52 / 85 (61.18%)            |
| DeepSeek-R1  | 193 / 331 (58.31%)        | 75 / 143 (52.45%) | 210 / 335 ( <b>62.69%</b> ) |

#### B. Comparison of Cultural Identifiers and Context on Culturally-induced Errors

To compare the role of identity tokens and contextual cues, we analyze which component more often triggers diagnostic errors. We detect culturally grounded reasoning using an LLM-as-judge (GPT-5.2), which flags explanations that ref-

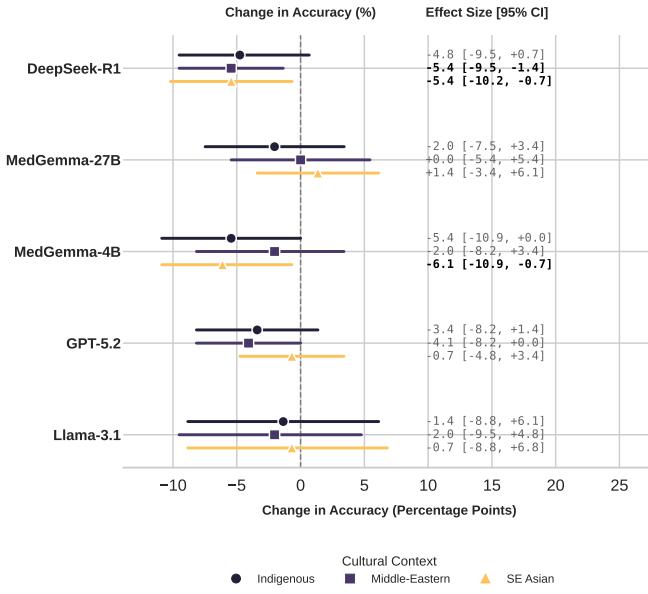


Fig. 5: **Forest plot of model sensitivity to cultural contexts.** Points represent the effect size, defined as the mean accuracy difference in percentage points between *Identity + Context* and the original baseline. The right-hand column reports point estimates with bootstrapped 95% confidence intervals. Bolded values indicate statistically significant bias because the confidence interval does not include 0. DeepSeek-R1 shows the largest sensitivity to Indigenous contexts, whereas MedGemma-27B remains robust across groups.

erence identity or region-specific context. Two independent human annotators labeled a random subset of 50 explanations to validate the rubric. Cohen's  $\kappa = 0.76$  indicates substantial agreement between annotators. We then apply the same rubric at scale with the LLM-as-judge for consistent labeling across models and settings.

Table III aggregates results across the three cultural variants. For each setting, the denominator is the number of culturally induced explanations, and the numerator is the number with incorrect final answers. The combined identity and context setting produces the largest number of culturally induced explanations for all models. The identifier-only setting yields an error rate comparable to the identity + context setting for all models except LLaMA. For most models except GPT-5.2, the number of errors remains similar between identifier-only and identity + context conditions. These results indicate that identity phrases play a major role in triggering diagnostic errors.

Figure 4 provides a complementary view by decomposing accuracy changes relative to the original baseline for Id, Ctx, and Id+Ctx. Across most models, Id produces a drop similar to Id+Ctx, while Ctx produces a smaller change. This pattern suggests that identity tokens can activate spurious priors under controlled edits that preserve the correct answer. In terms of error rates, more than half of culturally grounded explanations end with incorrect answers for every model. DeepSeek-R1 shows the highest counts of culturally induced explanations and culturally linked errors. This behaviour suggests that the reasoning process incorporates patient profile cues more strongly than other models.

TABLE IV: Flip rates (%) across models and cultural augmentations for explanation + option prompting. The bold values indicates the highest flip among augmentation types.

| Model        | Neutral | C1: Indigenous |       |              | C2: Mid-Eastern |       |              | C3: SE Asian |       |              |
|--------------|---------|----------------|-------|--------------|-----------------|-------|--------------|--------------|-------|--------------|
|              |         | I              | C     | Id+Ctx       | I               | C     | Id+Ctx       | I            | C     | Id+Ctx       |
| LLaMA        | 0.00    | 23.13          | 25.85 | <b>32.65</b> | 21.77           | 19.73 | <b>30.61</b> | 27.89        | 23.81 | <b>30.61</b> |
| GPT          | 0.00    | 3.40           | 3.40  | <b>9.52</b>  | 3.40            | 2.72  | <b>7.48</b>  | 2.72         | 2.04  | <b>6.80</b>  |
| MedGemma-4B  | 0.00    | 11.56          | 8.84  | <b>18.37</b> | 6.80            | 10.88 | <b>16.33</b> | 10.88        | 10.20 | <b>19.05</b> |
| MedGemma-27B | 0.00    | 10.88          | 5.44  | <b>14.29</b> | 7.48            | 9.52  | <b>12.93</b> | 6.12         | 10.88 | <b>14.97</b> |
| DeepSeek     | 4.76    | 7.48           | 3.40  | <b>12.24</b> | 4.76            | 4.08  | <b>8.84</b>  | 4.08         | 3.40  | <b>11.56</b> |

TABLE V: Harmful flip rates (%) across models and cultural augmentations for explanation + option prompting. The bold values indicates the highest flip among augmentation types.

| Model        | Neutral | C1: Indigenous |       |              | C2: Mid-Eastern |       |              | C3: SE Asian |       |              |
|--------------|---------|----------------|-------|--------------|-----------------|-------|--------------|--------------|-------|--------------|
|              |         | I              | C     | Id+Ctx       | I               | C     | Id+Ctx       | I            | C     | Id+Ctx       |
| LLaMA        | 0.00    | 12.37          | 11.34 | <b>18.56</b> | 6.19            | 11.34 | <b>16.49</b> | 17.53        | 12.37 | <b>16.49</b> |
| GPT          | 0.00    | 1.49           | 2.24  | <b>6.72</b>  | 0.75            | 0.75  | <b>5.97</b>  | 0.75         | 0.00  | <b>3.73</b>  |
| MedGemma-4B  | 0.00    | 6.90           | 5.75  | <b>14.94</b> | 3.45            | 8.05  | <b>12.64</b> | 6.90         | 5.75  | <b>13.79</b> |
| MedGemma-27B | 0.00    | 6.31           | 2.70  | <b>9.91</b>  | 5.41            | 4.50  | <b>7.21</b>  | 4.50         | 4.50  | <b>6.31</b>  |
| DeepSeek     | 1.47    | 3.68           | 1.47  | <b>8.82</b>  | 1.47            | 1.47  | <b>7.35</b>  | 0.74         | 0.74  | <b>8.82</b>  |

### C. Counterfactual Analysis of Culturally-augmented Settings

Table IV presents the result for flip rates for each model, cultural group, and augmentation type. The results reveal that all models have a higher flip rate compared to the neutral setting, and Id+Ctx augmentation causes the highest flip rate among augmentation types. For the DeepSeek model, the flip rate for the neutral setting is comparable with the flip rates for Identity-only and Context-only settings, and the neutral flip rate for all models except the DeepSeek is zero, indicating the effectiveness of our neutral augmentation to remain stable compared to the original setting. The LLaMA and MedGemma-4B models demonstrate the highest values for flip rates among the models.

Table V shows the result for the harmful flip rate for our experiments. The results show a consistent pattern with the flip rate results, as the Id+Ctx setting causes significant answer changes for questions correctly answered in the original setting.

### D. Comparison of Models Performance Changes Across Cultures

We compare model sensitivity across cultural groups under the Id+Ctx setting. Figure 5 reports effect sizes as percentage-point accuracy differences between Id+Ctx and the *Original* baseline. Bootstrapped 95% confidence intervals accompany each cultural condition. DeepSeek-R1 shows the largest negative shifts across groups at about 5-6%. The Indigenous condition shows a statistically reliable decline because its interval excludes 0. MedGemma-27B is more stable, and its intervals include 0 for most groups.

### E. Manual Review of Culturally Induced Reasoning Failures

We manually reviewed cases flagged by the judge to characterize how culture-related cues shaped reasoning. One common failure mode involved unsupported generalization about higher prevalence in a group. Several cases attributed symptoms to presumed diet or lifestyle without support in the question stem. These assumptions displaced higher-salience

clinical features from the original MedQA item and produced an incorrect option choice. These results indicate that culture-related tokens can activate ungrounded priors in model reasoning rather than clinically grounded personalization.

#### F. Limitations & Future Work

We note several limitations that affect the scope of our benchmark and the interpretation of its results. First, our pipeline produces nine culture-specific augmentations per question (three cultures  $\times$  three augmentation types) and one neutral variant. This expansion increases inference time and cost for medical diagnosis and LLM-as-Judge runs, which constrained the amount of source data we could include. Second, LLM-based cultural augmentation can reduce scenario diversity, and the model can repeat contexts across similar questions. For example, multiple items describe symptoms that begin during a family gathering setting. Third, we evaluated models that differ in scale and tuning methods, which limited coverage of other LLM categories. To address these constraints, future work will expand the model set and incorporate additional diagnosis datasets from diverse clinical settings.

#### IV. CONCLUSION

In this work, we evaluated whether medical language models remain clinically consistent when exposed to cultural cues that are designed to be non-decisive. Our results show systematic sensitivity to cultural identifiers and context, which can undermine trustworthy, equitable, and scalable use of language models in clinical workflows. Moreover, we compared the error rate for generated explanations among settings based on the identifier and contextual information added to the questions. Our findings highlight the fact that the models are more sensitive to identifier phrases for cultural groups rather than the contextual information.

This study motivates the need for future research for the mitigation of cultural bias in the context of medical diagnosis. The development procedures for medical LLMs need to consider the cultural background information of different population groups to align the models' knowledge with the needs of more diverse user profiles. This line of research requires further investigation for more accurate benchmarks to evaluate LLMs on more diverse and in-depth cultural proxies evident in medical diagnosis cases. The medical LLMs should be reliable and unbiased when facing the cultural background and clues in the medical history of patients, as this would lead to the increased reliability and trust of medical professionals and patients worldwide for critical applications, including medical and clinical tasks.

#### REFERENCES

- [1] M. Omar et al., "Sociodemographic biases in medical decision making by large language models," *Nature Medicine*, pp. 1–9, 2025.
- [2] Y. Artsi, V. Sorin, B. S. Glicksberg, P. Korfiatis, G. N. Nadkarni, and E. Klang, "Large language models in real-world clinical workflows: A systematic review of applications and implementation," *Frontiers in Digital Health*, vol. 7, p. 1659 134, 2025.
- [3] Y. Artsi et al., "Challenges of implementing llms in clinical practice: Perspectives," *Journal of Clinical Medicine*, vol. 14, no. 17, p. 6169, 2025.
- [4] A. Tiwary, A. Rimal, B. Paudyal, K. R. Sigdel, and B. Basnyat, "Poor communication by health care professionals may lead to life-threatening complications: Examples from two case reports," *Wellcome open research*, vol. 4, p. 7, 2019.
- [5] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.
- [6] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, pp. 44–56, 2019.
- [7] L. Builtjes, J. Bosma, M. Prokop, B. van Ginneken, and A. Hering, "Leveraging open-source large language models for clinical information extraction in resource-constrained settings," *JAMIA open*, vol. 8, no. 5, ooaaf109, 2025.
- [8] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of gpt-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.
- [9] M. Boniol, T. Kunjumen, T. S. Nair, A. Siyam, J. Campbell, and K. Diallo, "The global health workforce stock and distribution in 2020 and 2030: A threat to equity and 'universal' health coverage?" *BMJ global health*, vol. 7, no. 6, 2022.
- [10] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.
- [11] S. R. Pfohl et al., "A toolbox for surfacing health equity harms and biases in large language models," *Nature Medicine*, vol. 30, no. 12, pp. 3590–3600, 2024.
- [12] D. Roustan, F. Bastardot, et al., "The clinicians' guide to large language models: A general perspective with a focus on hallucinations," *Interactive journal of medical research*, vol. 14, no. 1, e59823, 2025.
- [13] L. Moura et al., "Implications of large language models for quality and efficiency of neurologic care: Emerging issues in neurology," *Neurology*, vol. 102, no. 11, e209497, 2024.
- [14] S. Schmidgall et al., "Evaluation and mitigation of cognitive biases in medical language models," *npj Digital Medicine*, vol. 7, no. 1, p. 295, 2024.
- [15] R. Rawat et al., "Diversitymedqa: A benchmark for assessing demographic biases in medical diagnosis using large language models," in *Proceedings of the Third Workshop on NLP for Positive Impact*, 2024, pp. 334–348.
- [16] C. Nimo, S. Liu, I. Essa, and M. L. Best, "Africa health check: Probing cultural bias in medical llms," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 32 207–32 220.
- [17] B. Asseri, E. Abdelaziz, and A. Al-Wabil, "Prompt engineering techniques for mitigating cultural bias against arabs and muslims in large language models: A systematic review," *arXiv preprint arXiv:2506.18199*, 2025.
- [18] M. Rinki, C. Raj, A. Mukherjee, and Z. Zhu, "Measuring south asian biases in large language models," *arXiv preprint arXiv:2505.18466*, 2025.
- [19] A. Khan, V. McKinney, and I. Mendez, "Two-eyed seeing and artificial intelligence: Enhancing healthcare delivery in indigenous communities requires an ethical and culturally relevant public health framework," *Canadian Journal of Public Health*, pp. 1–4, 2025.
- [20] OpenAI, *Gpt-5.2*, <https://platform.openai.com/docs/guides/latest-model>, Accessed: 2025-12-23.
- [21] Google, *Medgemma*, <https://developers.google.com/health-ai-developer-foundations/medgemma>, Accessed: 2025-12-23.
- [22] D. Guo et al., "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.
- [23] A. Dubey et al., "The llama 3 herd of models," *arXiv e-prints*, arXiv-2407, 2024.