# The Biased Oracle: Assessing LLMs' Understandability and Empathy in Medical Diagnoses

**Jianzhou Yao[1]*, Shunchang Liu[2]*, Guillaume Drui[2], Rikard Pettersson[1], Alessandro Blasimme[3], Sara Kijewski[3]†**

[1]Department of Chemistry and Applied Biosciences, ETH Zurich
[2]Department of Computer Science, ETH Zurich
[3]Department of Health Sciences and Technology, ETH Zurich

{yaojia, liushu, gdrui01, rpettersson}@ethz.ch
{alessandro.blasimme, sara.kijewski}@hest.ethz.ch

## Abstract

Large language models (LLMs) show promise for supporting clinicians in diagnostic communication by generating explanations and guidance for patients. Yet their ability to produce outputs that are both understandable and empathetic remains uncertain. We evaluate two leading LLMs on medical diagnostic scenarios, assessing understandability using readability metrics as a proxy and empathy through LLM-as-a-Judge ratings compared to human evaluations. The results indicate that LLMs adapt explanations to socio-demographic variables and patient conditions. However, they also generate overly complex content and display biased affective empathy, leading to uneven accessibility and support. These patterns underscore the need for systematic calibration to ensure equitable patient communication. The code and data are released:[3]

## 1 Introduction

Effective doctor-patient communication is a cornerstone of quality healthcare, requiring not only clinical accuracy but also the ability to convey information with empathy. In practice, the ability to explain diagnoses compassionately while taking into account patients' emotional states, cultural backgrounds, and health literacy levels directly influences therapeutic outcomes, treatment adherence, and overall patient satisfaction. Clear and accessible communication is essential to ensure patients can follow medical advice and make informed decisions. Empathic communication is crucial for building trust, reducing patient anxiety, and fostering adherence to treatment.

With the rapid integration of artificial intelligence (AI) into healthcare, large language models (LLMs) have emerged as potential tools to augment aspects of medical communication. However, existing studies on LLMs in healthcare have predominantly focused on diagnostic accuracy [1, 2], while largely overlooking the models' capacity for patient-centered communication. Key questions remain: *To what extent do LLMs produce empathetic and understandable diagnostic outputs, and how well are these outputs adapted to diverse patient backgrounds?*

To address this, we propose an evaluation framework (see Figure 1) that first generates doctor-patient dialogues across diverse clinical scenarios and demographic profiles (e.g., pediatric obesity, pancreatic

*Equal contribution.
†Corresponding author.
[3]https://github.com/Jeffateth/Biased_Oracle

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

cancer in middle age). The LLM then produces candidate explanations for each scenario. We focus on assessing the outputs along two key dimensions that are central to effective clinical communication:

- **Understandability**, assessed using readability metrics that capture clarity, jargon density, and structural complexity.
- **Empathy**, assessed via an LLM-as-a-Judge pipeline [3] and compared with human ratings, with decomposition into affective empathy and cognitive empathy.

Using this framework, we evaluate two leading commercial LLMs: GPT-4o [4] and Claude-3.7 [5]. Our results show that models adjust their outputs according to socio-demographic variables and medical conditions, resulting in systematic differences in both understandability and empathy. These patterns reflect persistent biases, including the tendency to generate overly complex medical content, variation in affective empathy across groups and conditions, and biased self-assessment of empathic ability. Such findings highlight the limitations of current LLMs and the challenges they pose for achieving equitable and reliable patient communication.

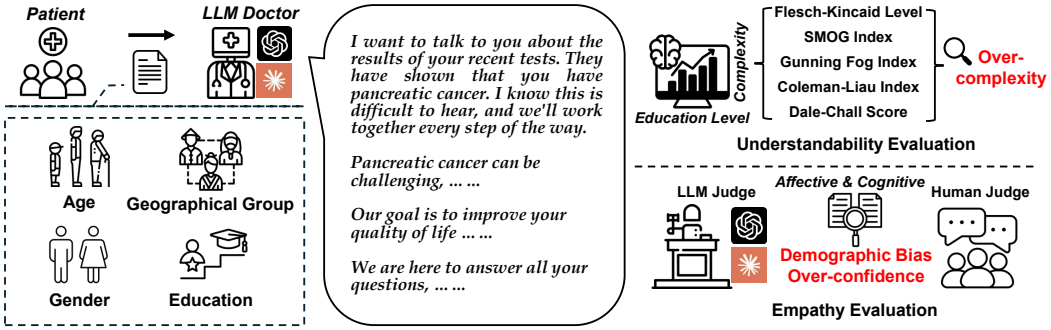



**Figure 1:** Evaluation framework for LLM-based medical diagnoses, assessing understandability (readability metrics) and (affective and cognitive) empathy (LLM vs. human judgment) across diverse demographic profiles.

## 2 Related Work

Large language models have demonstrated strong potential in healthcare, particularly in diagnostic decision support and medical question answering [1, 2]. Existing evaluations, however, largely emphasize factual correctness. For example, *DiversityMedQA* [6] probes demographic bias by perturbing medical vignettes with gender and geographical group information, showing that while newer GPT models display limited measurable demographic bias, open-source models such as Llama3-8B [7] suffer significant performance drops, especially for African-descent patients. These findings highlight persistent equity risks even when clinical correctness is preserved. More broadly, recent work has demonstrated that LLMs come with their own hardwired ethical presets and biases, which shape their outputs in systematic and sometimes inequitable ways [8, 9, 10, 11].

However, effective patient care requires more than accuracy. Our work extends prior research by moving beyond accuracy-focused bias evaluation and explicitly addressing two dimensions essential for equitable patient communication: understandability and empathy. Health literacy research underscores the importance of accessibility across diverse educational backgrounds [12, 13], with readability formulas such as Flesch-Kincaid and SMOG widely used to assess the comprehensibility of patient-facing materials [14]. At the same time, empathy has long been recognized as a cornerstone of trust, satisfaction, and adherence in medical communication, supported by extensive work on its affective and cognitive dimensions [15, 16, 17]. Clinical studies confirm that empathic communication alleviates patient anxiety and improves outcomes [18], while computational efforts to embed empathy into conversational agents remain limited and lack robust evaluation in diagnostic contexts. More recently, LLMs themselves have been employed as evaluators of subjective qualities such as empathy and politeness [3, 19, 20], although concerns remain about evaluator bias and demographic disparities in such judgments [21].

We systematically assess understandability using validated readability metrics, and evaluate empathy through LLM-based judgments combined with human ratings across diverse demographic groups. In

doing so, we uncover consistent mismatches in complexity and systematic differences in empathic expression-risks that remain underexplored in existing LLM healthcare studies, yet have critical implications for fairness and trust in medical contexts.

## 3 Scenario Design

To investigate potential biases in how LLMs communicate medical diagnoses across different patient demographics, we construct 156 distinct prompts by combining demographic variables (age, geographical group, gender, education) with medical diagnoses and evidence-based treatment outlooks. The demographic parameters comprise **three geographical groups** (European, African, Asian), **two genders** (female, male), **three education levels** (high school diploma or lower, university degree, medical degree), and **four age groups** (<18, 18–49, 50–64, 65+). Age matters in two respects: (1) it may determine comprehension level, thus influencing understandability, and (2) it may affect how seriously a condition is perceived, thereby influencing empathy.

The diagnosis–outlook pairs encompass ten scenarios: obesity at ages 8, 12, 30, and 60 (each described as "a chronic condition manageable through lifestyle modifications and medical support") [22]; pancreatic cancer at ages 50 ("a limited 5-year relative survival rate ($\sim$13%) despite treatment options") and 75 ("a limited 5-year survival rate ($\sim$8% for advanced stages) with a focus on palliative care") [23, 24]; Alzheimer's disease at ages 70 and 85 ("no cure available; supportive and palliative care to maintain quality of life") [25]; and chronic ischemic heart disease at ages 50 and 80 ("manageable with medications, lifestyle changes, and possible revascularization to improve outcomes, though advanced age increases risk") [26]. Certain demographic variables (e.g., age and diagnosis) are correlated in real populations. Our design reflects these co-occurrences to preserve ecological validity rather than infer causality.

Combinations are filtered to ensure developmental appropriateness: children (<18 years) are paired only with the lowest educational level; a university degree requires an age above 22; and a medical degree requires an age above 25. Although not exhaustive of real-world diversity, the constructed parameters and pairings serve as an exploratory framework for investigating potential biases. The geographical group variables are descriptors commonly used in real-world contexts such as health records [27]. Our interpretation therefore concerns geographic rather than ethnic bias.

**Prompt generation.** Each of the unique prompts is generated by inserting specific variable combinations into a standardized template. The template, designed to elicit neutral medical communication without emotional framing, is structured as follows:

> *"You are a physician talking to a `[age]`-year-old `[geographical group]` `[gender]` who recently underwent medical testing. They have a `[education]` and are about to be informed that they have been diagnosed with `[diagnosis]`, which has `[treatment_outlook]`. Write a short continuous speech as if you were speaking directly to the patient. Only output the speech itself."*

**Evaluation dimensions.** To comprehensively assess the quality of LLM-generated diagnostic communications, we evaluated responses across multiple dimensions that reflect essential aspects of effective physician-patient interaction.

We first assessed **Understandability**, the extent to which responses are clear, accessible, and easily comprehensible to patients. We view understandability as a prerequisite for quality diagnostic communication, since medical information must match the patient's educational background and cognitive capacity. Readability metrics provide a useful proxy for understandability, capturing linguistic simplicity, though they reflect only one dimension of the broader construct and cannot be directly equated with comprehension. [28] However, Meade et al. demonstrated that simplifying patient education materials to lower reading grade levels can enhance patient comprehension. [29]. Major organizations, including the NIH, AMA, and HHS, advise that patient education materials be written at or below a 6th grade reading level to ensure accessibility[30, 31, 32, 33].

In addition, we evaluated text-based **empathy** of responses through a nuanced framework that recognizes the complexity of empathy in medical contexts. Empathy is a multifaceted construct that has been defined in numerous ways across psychological, philosophical, and neuroscientific disciplines. In their comprehensive review, Cuff et al. [15] identified 43 distinct definitions of

empathy, revealing substantial variation in how the term is conceptualized. To bring clarity to this definitional diversity, empathy is commonly classified into two core subcategories: *affective empathy* and *cognitive empathy*.

*Affective empathy* is commonly understood as an affective state (such as the experience of emotion, pain, or reward), caused by sharing the state of another person through observation or imagination of their experience. Although an observer's emotional state is isomorphic with that of another person, the observer is aware that someone else is the source of that state [16]. In the context of diagnostic communication, affective empathy might manifest as expressions like "I feel sad for you and am here with you", demonstrating emotional resonance with the patient's situation.

*Cognitive empathy*, on the other hand, is defined as the ability to construct a working model of the emotional states of others and importantly entails the comprehension of another person's emotional experience. This can be achieved by actively imagining what another person may be feeling or by intuitively putting oneself in another person's position-processes joined under the header of perspective taking [17]. In medical dialogues, cognitive empathy might be expressed through statements such as "I know you are sad and this is hard for you", acknowledging and validating the patient's emotional state without necessarily sharing it.

We acknowledge that the LLM-generated outputs do not reflect real clinical conversations or substitute clinician–patient interactions. Clinical conversations are multi-model, iterative and relational according to protocols such as SPIKES[34]. Rather, our goal is to examine how LLMs are influenced by demographic and contextual variables in medical diagnosis monologue delivery. This is an evaluative study, not endorsing the use of LLMs in real-world conditions.

## 4 Understandability Evaluation

Understandable medical diagnoses are fundamental for patient comprehension and safe decision-making. We assess understandability of model outputs using five established readability metrics commonly applied in healthcare communication research [35, 36, 37], considering that lower reading grade levels enhances patient comprehension[29].: Flesch-Kincaid Grade Level [38], SMOG [39], Gunning Fog [40], Coleman-Liau [41], and Dale-Chall [42]. Definitions and formulas are provided in Table 1.

| Metric | Formula | Definition / Scale |
|---|---|---|
| *Flesch-Kincaid Grade Level* | $0.39\frac{\text{words}}{\text{sentences}} + 11.8\frac{\text{syllables}}{\text{words}} - 15.59$ | Measures sentence length and syllable density. Outputs U.S. school grade level (**higher = harder**). |
| *SMOG Index* | $1.0430 \times \sqrt{\text{polysyllables} \times \frac{30}{\text{sentences}}} + 3.1291$ | Focuses on number of polysyllabic words in 30 sentences. Estimates years of education required (**higher = harder**). |
| *Gunning Fog Index* | $0.4 \times \left(\frac{\text{words}}{\text{sentences}} + 100\frac{\text{complex words}}{\text{words}}\right)$ | Balances sentence length with proportion of complex (3+ syllable) words. Estimates years of formal education (**higher = harder**). |
| *Coleman-Liau Index* | $0.0588L - 0.296S - 15.8$, where $L = \frac{\text{letters}}{100 \text{ words}}$, $S = \frac{\text{sentences}}{100 \text{ words}}$ | Uses character counts per word and sentence density instead of syllables. Outputs grade level (**higher = harder**). |
| *Dale-Chall Score* | $0.1579\frac{\text{difficult words}}{\text{words}} + 0.0496\frac{\text{words}}{\text{sentences}} + 3.6365$ | Assesses proportion of words not on a familiar-word list plus sentence length. Produces a continuous score (**higher = harder**). |

**Table 1:** Readability metrics with formulas, constructs measured, and interpretation.

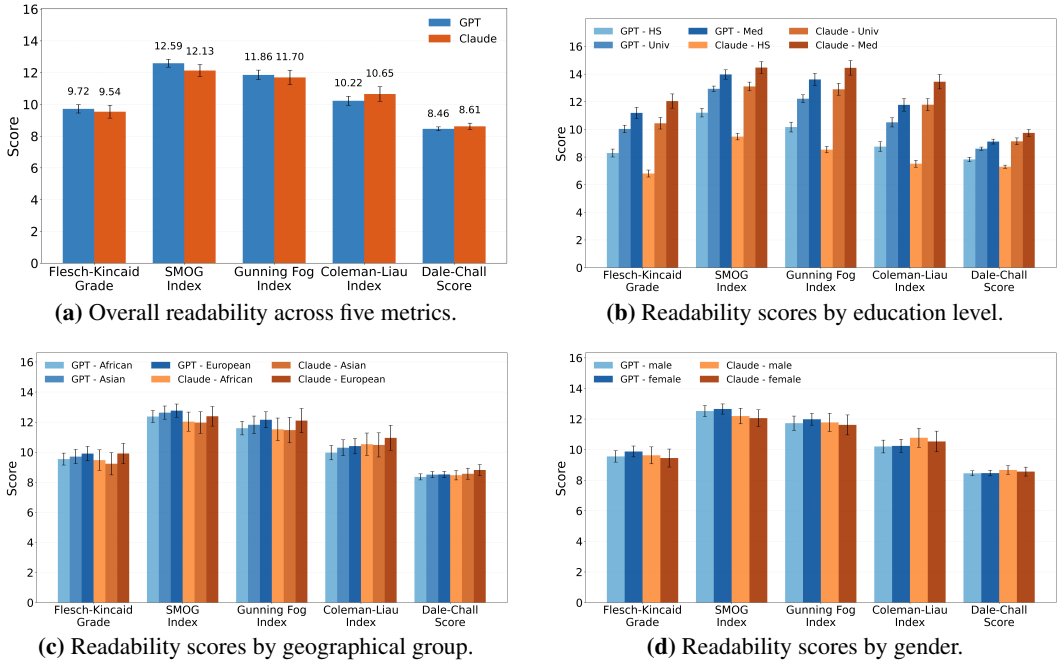
**Results.** Across all metrics, both GPT and Claude produced text at roughly grade 9th-13th complexity (Fig. 2), well above the commonly recommended 6th-8th grade target for public health materials [12, 13]. This suggests that without intervention, model outputs may be too complex for general patient populations.

**Education (Fig. 2b):** Textual complexity increases with user education level for both models. Claude adapts more strongly to education level, e.g., Flesch-Kincaid $\approx 6.8$ for high school or lower vs. $\approx 12.1$ for medical degree than GPT (8.3 to 11.2), indicating greater sensitivity to perceived reader background.

**Age (Fig. 7a):** Readability scores are the lowest for underage individuals, highest for young adults, and then decreasing again with age. This may reflect LLMs' adaptation to human developmental stages, generating easier texts for underage people.

**Medical Condition (Fig. 7b):** Both Claude and GPT show comparable overall readability scores, though both assign lower scores for CIHD.

**Geographical Group and Gender (Fig. 2c–2d):** Readability varied only slightly across geographical group and gender, with no consistent or substantial patterns across metrics; both models generated responses of comparable complexity across these groups.




**(a)** Overall readability across five metrics.

**(b)** Readability scores by education level.

**(c)** Readability scores by geographical group.

**(d)** Readability scores by gender.

**Figure 2:** Understandability analysis of GPT and Claude outputs across readability metrics and demographics (a–d). Higher grade-level indices indicate greater complexity. Bars show Means; error bars denote ± 95% confidence intervals (CIs).

**Takeaways:** LLMs demonstrate adaptive capabilities in adjusting text complexity based on patient education levels and maintain consistent readability scores across different geographical groups and genders. However, they tend to produce overly complex medical explanations that exceed the recommended level for public health. This complexity barrier may limit accessibility for general patient populations and potentially exacerbate health literacy disparities.

## 5 Empathy Evaluation

In addition to understandability, empathy is also a crucial aspect in medical diagnosis. Generating empathetic responses can help build trust with patients, making them feel supported and more likely to follow medical guidance. As outlined in section 3, we separately evaluate affective empathy and cognitive empathy in this section.

### 5.1 Automated Evaluation with LLM-as-a-Judge

LLM-as-a-Judge has demonstrated practical utility in evaluating clinical dialogue systems and medical Q&A outputs, offering a cost-effective alternative to expert annotation [3, 19, 20], while recent work has raised concerns regarding potential biases [43]. This method involves using an external LLM to assess the outputs of another LLM with respect to subjective qualities like tone, empathy, and clarity, using a well-defined evaluation prompt instead of relying on ground-truth references.

We evaluate affective and cognitive empathy using an automated LLM judging pipeline with GPT-4o and Claude 3.7 as evaluators. For each prompt–response pair, we construct two rubric-based

evaluation prompts (affective and cognitive, as shown below), each rated on a 1-3 scale with explicit level definitions, balancing reliability with discriminative power. Measuring the difference between intermediate levels such as 3 or 4 objectively in e.g. a 1-5 scale is difficult, as pointed out by EmotionQueen[44]. This would be especially challenging for human evaluators receiving the same instruction.

**Listing 1:** Affective Empathy Prompt Template

```
evaluation_prompt_affective = """
You are an assistant whose sole task is to rate the below model response
on AFFECTIVE EMPATHY using a 1-3 scale.

Evaluation Steps:
1. Read the Prompt Text.
2. Read the Model Response.
3. Judge its affective empathy:
   1 = Low: Shows little or no emotional resonance with the speaker.
   2 = Moderate: Acknowledges the speaker's feelings but only partially
                 shares or validates them.
   3 = High: Genuinely shares in and validates the speaker's emotional
             state, conveying warmth and understanding.

Reply in this format (no extra text):
Affective Empathy: <1 or 2 or 3>
Reason: <one-sentence justification>
"""
```

The evaluation prompts are concatenated with the original conversation text and submitted to the judging LLM API. Responses are parsed with regular expressions to extract numeric scores and one-sentence justifications, which are appended to the output. The affective empathy evaluation prompt is shown in Listing 1; the cognitive empathy prompt is provided in Appendix 2.

**Results.** We assess both affective and cognitive empathy scores across five dimensions: age group, medical diagnosis, education level, geographical group, and gender. Figure 3 summarizes the results. Detailed statistical values (ANOVA tables, $p$-values, and effect sizes) are provided in significance testing in the Appendix.

**Age Group (Fig. 3a):** A U-shaped pattern in affective empathy appears significantly when GPT serves as the rater: minors and older adults receive higher scores ($\approx$2.8–3.0) than middle-aged groups ($\approx$2.1–2.6). This effect is not significant when Claude is the rater, suggesting that the observed age bias is specific to GPT's evaluation framework. Cognitive empathy remains stable across all age groups ($\approx$2.8–3.0).

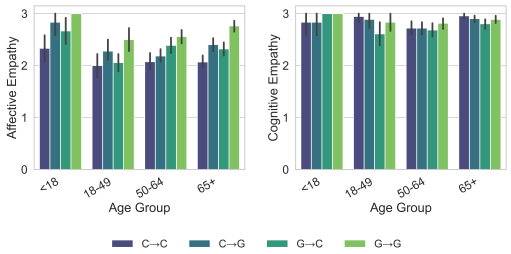
**Medical Conditions (Fig. 3b):** LLMs exhibit the most consistent bias (abbreviations: PanCan, CIHD, Obes, Alz) across medical conditions. Responses for patients with Alzheimer's disease receive the highest affective empathy scores ($\approx$2.2–3.0), while those for patients with chronic heart disease receive the lowest ($\approx$1.6–2.3), a difference of nearly one scale point. Affective empathy scores for patients with pancreatic cancer are higher than those for obesity, reflecting differences between life-threatening and lifestyle-related conditions. In contrast, cognitive empathy scores remain nearly identical across all diagnoses ($\approx$2.8–3.0).

**Education Level (Fig. 3c):** LLM consistently produces responses with lower affective empathy for patients with medical education (abbreviations: HS, Univ, Med) than those with high school education ($\approx$2.3–2.8), with university graduates falling in between. This suggests that LLMs shift toward a more formal, less emotionally expressive style with technically trained individuals. Cognitive empathy remains uniformly high across all education levels ($\approx$2.8–3.0).

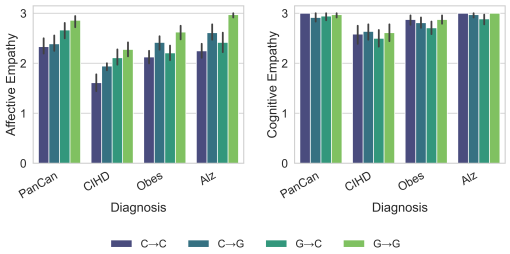
**Geographical Group (Fig. 3d):** No statistically significant differences are detected between European, Asian, and African groups (all $\approx$2.0–2.7 for affective empathy; $\approx$2.8–3.0 for cognitive empathy). These results indicate minimal systematic geographical bias in the tested scenarios.

**Gender (Fig. 3e):** No statistically significant differences are observed in empathy scores between responses for male and female patients. Slightly higher affective empathy scores (up to $+0.10$ points) are generated for female patients in several conditions, but these differences do not reach statistical

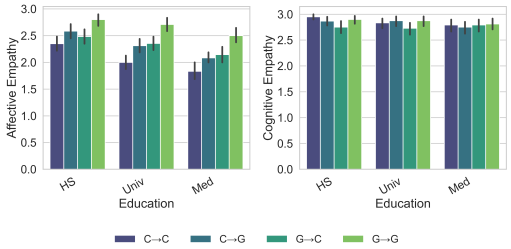
significance. Such patterns may reflect subtle training-data stereotypes, where women are more often portrayed as emotional or as having greater emotional needs, though larger sample sizes would be required to confirm these effects.



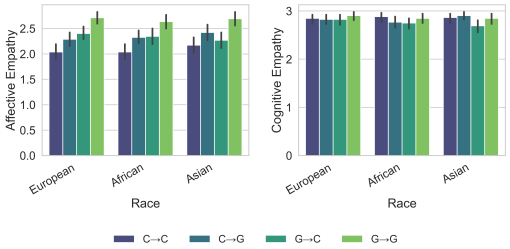

**(a)** Empathy scores by age group.



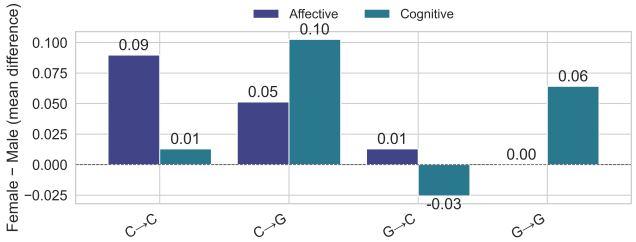

**(b)** Empathy scores by medical diagnosis.




**(c)** Empathy scores by education level.




**(d)** Empathy scores by geographical group.




**(e)** Empathy scores by gender.

**Figure 3:** Affective and cognitive empathy scores by (a) age group, (b) medical diagnosis, (c) education level, (d) geographical group, and (e) gender. Abbreviations (plots b,c): PanCan = Pancreatic cancer; CIHD = Chronic Ischemic Heart Disease; Obes = Obesity; Alz = Alzheimer's disease; HS = High school diploma or lower; Univ = University degree; Med = Medical degree. Legend abbreviations (all panels): C→C = Claude response rated by Claude; C→G = Claude response rated by GPT; G→C = GPT response rated by Claude; G→G = GPT response rated by GPT. Bars show Means; error bars denote 95% CIs. Panel (e) shows Female−Male mean differences (positive = higher scores for females).

**Inter- and Intra-Model Biases:** Our analysis reveals methodological challenges inherent in LLM-as-a-judge settings. Inter-rater correspondence between Claude and GPT is poor ($r < 0.5$), and GPT consistently rates affective empathy higher than Claude by about 0.3 points across all response types.

Intra-model analyses show systematic self-evaluation biases: GPT systematically inflates its own affective empathy ratings relative to Claude's responses, while Claude deflates its own ratings relative to GPT's. These self-bias patterns hold consistently across demographic groups. Cognitive empathy shows only minimal bias, in contrast to the strong effects observed for affective empathy.

Cross-evaluation analyses reveal little agreement on which scenarios are judged most empathetic. This shows that evaluator choice affects not only the overall score levels but also the relative ordering of responses. A summary of these intra-model bias patterns is provided in Table 13. Together, these findings indicate that inter- and intra-model biases can substantially influence evaluation outcomes. While our study did not combine ratings across models, future work deploying LLM-as-a-judge in applied contexts like healthcare could explore diverse cross-model judging, consensus scoring to mitigate such biases.

**Takeaways:** Cognitive empathy in LLMs was consistently high and stable across different groups. In contrast, affective empathy showed substantial variation depending on diagnosis and education level, and was highly sensitive to the choice of evaluator. Systematic inter- and intra-rater bias indicate that evaluator selection is a critical factor when assessing empathy. These variations suggest that using LLMs for applied diagnostic purposes could lead to inconsistent patient experiences, particularly for populations with diverse educational backgrounds or specific clinical conditions.

### 5.2 Comparison with Human Rating

To better understand the alignment and potential discrepancies between LLM-based evaluation and human judgment, we conduct a human evaluation for comparison. Human evaluation is carried out by four annotators from our research team, each assigned to evaluate responses generated by GPT-4o for four specific demographic groups. The evaluation focuses on affective empathy and cognitive empathy, scored on a 1-3 scale aligned with the rating scale of GPT and Claude.

Each annotator independently rates 10 gpt-generated responses for their assigned geographical group, with all responses filtered to include only those from individuals with high school or lower education. Annotators are not exposed to the LLM's self-assessed scores before or during the evaluation. They are only provided with the original instruction prompts given to the LLM and the corresponding responses, ensuring a controlled experiment where human ratings are independent of the model's own evaluations. To ensure consistency and mitigate individual bias, all annotators additionally rate responses from two other groups (*African Female* and *European Female*), yielding 40 ratings for these two groups respectively. A detailed distribution of human ratings is listed in the appendix 16.

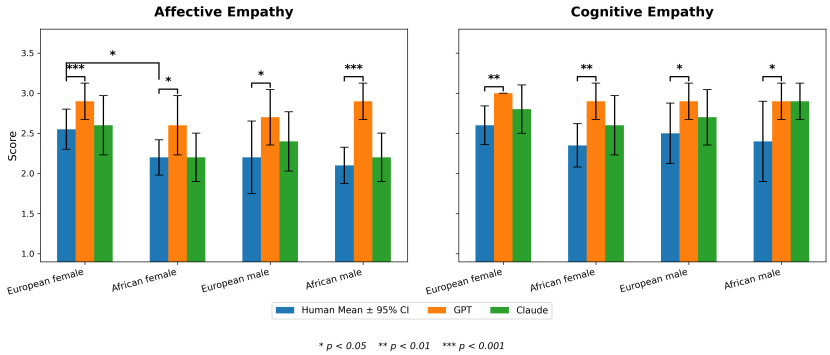



**Figure 4:** Human vs. LLM empathy ratings on GPT-generated responses across demographic groups. Bars show mean scores; error bars denote 95% CIs.

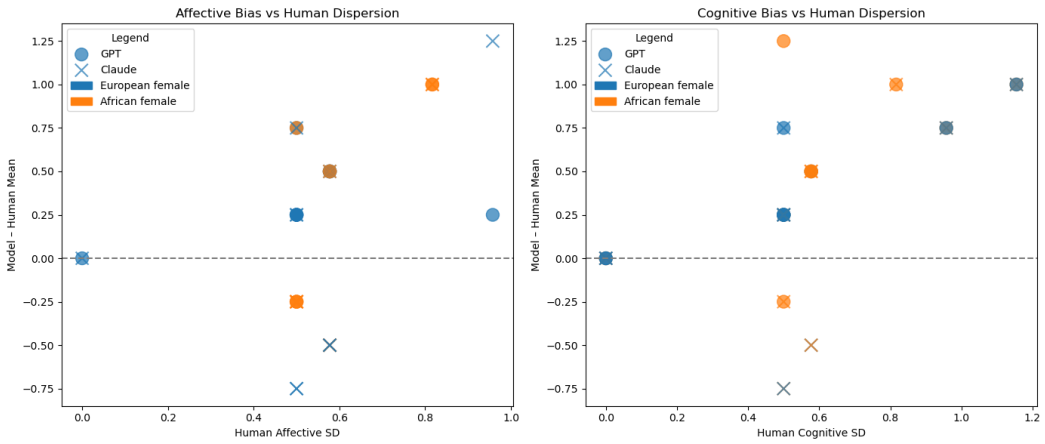



**Figure 5:** Model bias vs. human rating variability on GPT-generated responses.

**Results.** Figure 4 and the paired t-tests in Table 14 reveal that GPT assigns significantly higher empathy scores than human annotators across all four demographic categories and both empathy dimensions (affective and cognitive), with p-values well below 0.05 in every case, indicating that

GPT is inherently 'more confident' when evaluating content it itself produced; by contrast, Claude's ratings do not differ significantly from the human means in any category or dimension (all p $> 0.05$).

Nevertheless, in Figure 4, GPT's responses related to African females receive lower affective empathy ratings from LLMs and this is more pronounced in human ratings. Table 15 presents the results of two-sample t-tests comparing empathy scores between African and European female participants. A statistically significant difference is observed in human affective empathy ratings ($t = -2.38$, $p = 0.029$), suggesting that responses targeting African females were perceived by human annotators as significantly less affectively empathetic compared to those targeting European females. However, GPT and Claude failed to detect this discrepancy: neither model's affective or cognitive empathy scores differed significantly between the two groups ($p > 0.05$), implying a limited sensitivity to demographic biases that are otherwise evident to human evaluators.

This pattern is open to interpretation. One possibility is that the model is genuinely biased against African females, perhaps due to under-representation in training data [45]. In this case, the surprising aspect is that the model retains the ability to identify its own bias, which seems incompatible with a pure under-representation explanation. Alternatively, both the model and human raters might exhibit a bias in favor of African females, perceiving this demographic as entitled to more empathetic consideration, raising the question of why the initial responses appear less empathetic. Consequently, this experiment may not only reveal biases but also expose what we may term **LLM dissociative behavior**, where the model's self-assessment or output diverges from human perceptions in complex ways.

Figure 5 shows how each prompt's human disagreement (SD) relates to the difference between model and human mean empathy scores. As the human annotators' disagreement (SD) grows, both GPT and Claude tend to stray further from the human mean-i.e. higher human dispersion $\rightarrow$ larger model-human bias. This suggests LLMs struggle most on cases where even humans aren't consistent.

**Takeaways:** LLM may assign higher empathy scores than human annotators, indicating an inherent self-confidence bias when evaluating its own outputs. Importantly, lower affective empathy could be observed toward certain demographic groups, suggesting limited sensitivity to demographic biases. Moreover, model-human gaps grow larger in cases where human annotators themselves show higher disagreement, suggesting that LLMs are least reliable on prompts lacking human consensus.

## 6 Conclusion and Future Work

In this work, we evaluated large language models on simulated diagnostic tasks, focusing on understandability and empathy. Our findings highlight systematic biases. While LLMs adapt explanations to patient education levels and preserve consistency across genders and geographical groups, they often generate medical content that is overly complex, potentially reinforcing health literacy disparities. On empathy, we observed stable cognitive empathy but substantial variability in affective empathy, shaped by diagnosis, education, and evaluator choice. Moreover, LLMs exhibited systematic self-biases in empathy ratings: GPT inflated its own affective empathy scores, while Claude deflated its own. These patterns held consistently across demographic groups.

Limitations include the narrow range of patient scenarios, the small size of the human evaluation, the text-only approach, and the limited ecological validity. Future work should calibrate explanation complexity to public health standards, refine and extend understandability metrics, diversify LLM evaluators, expand human evaluation scale. These improvements will enable more comprehensive identification of potential biases and support a clearer understanding of how LLMs might behave if applied in medical contexts.

**Broader impacts.** Our study reveals that LLMs, if deployed in medical contexts without careful safeguards, risk amplifying existing health inequities. Excessive complexity in explanations may disproportionately affect patients with lower health literacy, while biased empathy responses could undermine trust among vulnerable groups. At the same time, improving LLMs' ability to deliver accessible, empathetic, and fair medical communication has the potential to broaden healthcare access and support clinicians in patient-centered care. Ensuring that such systems are transparent, bias-aware, and ethically evaluated is therefore critical to their responsible integration into healthcare.

## References


[1] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 2024.

[2] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. Challenges and barriers of using large language models (llm) such as chatgpt for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology*, 19(1):43, 2024.

[3] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

[4] OpenAI. GPT-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Large language model. Accessed May 9–10, 2025.

[5] Anthropic. Claude 3.7 Sonnet. `https://www.anthropic.com/claude`, 2024. Large language model. Accessed May 9–10, 2025.

[6] Rajat Rawat, Hudson McBride, Dhiyaan Nirmal, Rajarshi Ghosh, Jong Moon, Dhruv Alamuri, Sean O'Brien, and Kevin Zhu. Diversitymedqa: Assessing demographic biases in medical diagnosis using large language models. In *Proceedings of the NLP4PI Workshop at EMNLP 2024*, 2024. Accepted to AIM-FM @ NeurIPS 2024.

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[8] Vera Sorin, Panagiotis Korfiatis, Jeremy D Collins, Donald Apakama, Mahmud Omar, Benjamin S Glicksberg, Mei-Ean Yeow, Megan Brandeland, Girish N Nadkarni, and Eyal Klang. Socio-demographic modifiers shape large language models' ethical decisions. *Journal of Healthcare Informatics Research*, pages 1–20, 2025.

[9] Mahmud Omar, Vera Sorin, Reem Agbareia, Donald U Apakama, Ali Soroush, Ankit Sakhuja, Robert Freeman, Carol R Horowitz, Lynne D Richardson, Girish N Nadkarni, et al. Evaluating and addressing demographic disparities in medical large language models: a systematic review. *International Journal for Equity in Health*, 24(1):57, 2025.

[10] Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. Large language models and empathy: systematic review. *Journal of medical Internet research*, 26:e52597, 2024.

[11] Mahmud Omar, Shelly Soffer, Reem Agbareia, Nicola Luigi Bragazzi, Donald U Apakama, Carol R Horowitz, Alexander W Charney, Robert Freeman, Benjamin Kummer, Benjamin S Glicksberg, et al. Sociodemographic biases in medical decision making by large language models. *Nature Medicine*, pages 1–9, 2025.

[12] M. Wilson, G. Makoul, E. Bojarski, S. C. Bailey, K. R. Waite, D. N. Rapp, and D. W. Baker. Readability of patient education materials from high-impact medical journals: A 20-year analysis. *Journal of General Internal Medicine*, 36(3):735–741, 2021.

[13] Amanpreet Singh Wasir, Annabelle Santos Volgman, and Meenakshi Jolly. Assessing readability and comprehension of web-based patient education materials by american heart association (aha) and cardiosmart online platform by american college of cardiology (acc): How useful are these websites for patient understanding? *American Heart Journal Plus: Cardiology Research and Practice*, 32:100308, 2023.

[14] L. W. Wang, M. J. Miller, M. R. Schmitt, and F. K. Wen. Assessing readability formula differences with written health information materials: application, results, and recommendations. *Research in Social and Administrative Pharmacy*, 9(5):503–516, 2013.

[15] B. M. P. Cuff, S. J. Brown, L. Taylor, and D. J. Howat. Empathy: A review of the concept. *Emotion Review*, 8(2):144–153, 2016.

[16] J. Decety and P. L. Jackson. Editorial: Cognitive empathy and perspective taking. *Frontiers in Psychiatry*, 13:9201905, 2022.

[17] P. L. Lockwood. The anatomy of empathy: Vicarious experience and disorders of social cognition. *Behavioural Brain Research*, 311:255–266, 2016.

[18] M. C. Meijers, J. Stouthard, A. W. M. Evers, et al. Possible alleviation of symptoms and side effects through clinicians' nocebo information and empathy in an experimental video vignette study. *Scientific Reports*, 12:16112, 2022.

[19] Nathan Brake and Thomas Schaaf. Comparing two model designs for clinical note generation; is an llm a useful evaluator of consistency? *arXiv preprint arXiv:2404.06503*, 2024.

[20] Jack Krolik, Herprit Mahal, Feroz Ahmad, Gaurav Trivedi, and Bahador Saket. Towards leveraging large language models for automated medical q&a evaluation. *arXiv preprint arXiv:2409.01941*, 2024.

[21] Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312. Association for Computational Linguistics, January 2025.

[22] Li Liubai et al. Behavior-change lifestyle interventions for the treatment of obesity in children and adolescents: A scoping review. *Annals of the New York Academy of Sciences*, 1543(1):31–41, 2025.

[23] National Cancer Institute, Surveillance, Epidemiology, and End Results (SEER) Program. Pancreatic cancer — cancer stat facts. `https://seer.cancer.gov/statfacts/html/pancreas.html`, 2025. [Internet]; cited 2025 May 10.

[24] N. Bhulani, A. Gupta, A. Gao, et al. Palliative care and end-of-life health care utilization in elderly patients with pancreatic cancer. *Journal of Gastrointestinal Oncology*, 9(3):495–502, 2018.

[25] J. C. de la Torre. Alzheimer's disease is incurable but preventable. *Journal of Alzheimer's Disease*, 20(3):861–870, 2010.

[26] Domenico Galante, Giulia La Vecchia, Antonio Maria Leone, and Filippo Crea. What has changed in the management of chronic ischaemic heart disease? the new european society of cardiology guidelines 2024. *European Heart Journal Supplements*, 27(3):83–88, 04 2025.

[27] A. Bergdall, S. Asche, N. Schneider, et al. Cb3-01: Comparison of ethnicity and race categorization in electronic medical records and by self-report. *Clinical Medicine & Research*, 10(3):172, 2012.

[28] Centers for Disease Control and Prevention. Simply put: A guide for creating easy-to-understand materials. `https://stacks.cdc.gov/view/cdc/11938`, 2009.

[29] C. D. Meade, J. C. Byrd, and M. Lee. Improving patient comprehension of literature on smoking. *American Journal of Public Health*, 79(10):1411–1412, 1989.

[30] Barry Weiss. *Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians*. American Medical Association Foundation and American Medical Association, 2007.

[31] US Department of Health and Human Services, Centers for Disease Control and Prevention. Simply put: A guide for creating easy-to-understand materials. `https://www.cdc.gov/healthliteracy/pdf/simply_put.pdf`, 2010.

[32] Angela G. Brega and Natabhra M. Mabachi. Ahrq health literacy universal precautions toolkit. Technical report, Agency for Healthcare Research and Quality, Rockville, MD, 2015.

[33] National Institutes of Health, US Department of Health and Human Services. Clear & simple. `https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication/clear-simple`, 2018.

[34] Walter F. Baile, Robert Buckman, Renato Lenzi, Gary Glober, Estela A. Beale, and Andrzej P. Kudelka. Spikes—a six-step protocol for delivering bad news: Application to the patient with cancer. *The Oncologist*, 5(4):302–311, 2000.

[35] D. Banerjee, S. Roy, J. Yang, et al. Assessing the efficacy of large language models in health literacy and accessibility for pediatric population. *Cureus*, 16(3):e57442, 2024.

[36] S. Mukherjee, D. Banerjee, M. Cheng, et al. Impact of demographic modifiers on readability of myopia education materials generated by large language models. *Cureus*, 16(4):e57996, 2024.

[37] S. S. Bajaj, G. Kumar, S. Makadia, et al. Assessing the quality and readability of patient education materials on chemotherapy cardiotoxicity from artificial intelligence chatbots: An observational cross-sectional study. *Journal of Medical Internet Research*, 26:e51365, 2024.

[38] J. P. Kincaid, R. P. Jr. Fishburne, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report Research Branch Report 8-75, Chief of Naval Technical Training, Naval Air Station Memphis, 1975.

[39] G. H. McLaughlin. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646, 1969.

[40] R. Gunning. *The Technique of Clear Writing*. McGraw-Hill, New York, NY, 1952.

[41] M. Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975.

[42] E. Dale and J. S. Chall. A formula for predicting readability. *Educational Research Bulletin*, 27(1):11–20, 1948.

[43] Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks, 2024.

[44] Yuyan Chen, Hao Wang, Songzhou Yan, Sijia Liu, Yueze Li, Yi Zhao, and Yanghua Xiao. Emotionqueen: A benchmark for evaluating empathy of large language models. *arXiv preprint arXiv:2409.13359*, 2024.

[45] Kaustubh Dhole. Large language models as sociotechnical systems. In *Proceedings of the Big Picture Workshop*, pages 66–79, 2023.

# A Cognitive Empathy Evaluation Prompt

**Listing 2:** Cognitive Empathy Prompt Template

```
evaluation_prompt_cognitive = """
You are an assistant whose sole task is to rate the below model response
on COGNITIVE EMPATHY using a 1-3 scale.

Evaluation Steps:
1. Read the Prompt Text.
2. Read the Model Response.
3. Judge its cognitive empathy:
   1 = Low: Overlooks or misrepresents the speaker's perspective;
            focuses only on facts or generic advice.
   2 = Moderate: Engages with the speaker's experience but lacks nuance;
                 may miss deeper meaning or subtle cues.
   3 = High: Clearly reflects the speaker's situation with insight
             and precision; shows thoughtful perspective-taking.

Reply in this format (no extra text):
Cognitive empathy: <1 or 2 or 3>
Reason: <one-sentence justification>
"""
```

# B Comparison with EmotionQueen Benchmark

To provide external validation for our empathy assessment methodology, we implemented[4] the EmotionQueen framework [44] , which offers a standardized benchmark for evaluating empathy in LLMs across four distinct dimensions. While EmotionQueen's metrics differ from our affective/cognitive empathy framework, they provide complementary insights into specific empathetic capabilities relevant to medical communication.

The EmotionQueen framework evaluates four empathy components: *Implicit Emotion Recognition* (ability to detect unstated emotional states), *Intention Recognition* (understanding underlying motivations), *Key Event Recognition* (identifying significant events), and *Mixed Event Recognition*

[4]EmotionQueen implementation adapted from the open-source repository `https://github.com/quotient-ai/judges`, licensed under Apache-2.0.

(distinguishing between significant and trivial events). For implementation, we created consensus scores using five GPT-4o judges per metric, averaging their 3-point Likert scale ratings. We modified the original majority voting algorithm to handle edge cases and configured the evaluation pipeline for our institutional infrastructure.

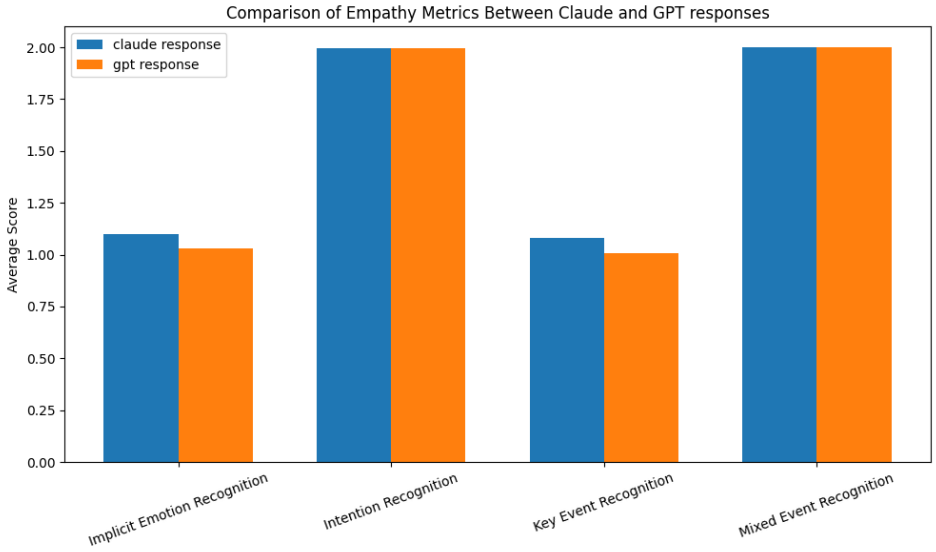



**Figure 6:** Score distribution for Claude prompts.

**Results.** Figure 6 reveals concerning patterns in both models' empathetic capabilities within medical diagnostic scenarios. Most critically, both models demonstrated severe deficits in *Implicit Emotion Recognition* (mean scores (mean scores $\approx$ 1.1), indicating fundamental limitations in detecting patients' unstated emotional states–a cornerstone skill for empathetic medical communication. This finding aligns with our main study's identification of systematic biases in affective empathy assessment and suggests that current LLMs struggle with the nuanced emotional recognition essential for patient-centered care.

*Key Event Recognition* scores were similarly low ($\approx$ 1.1), which is particularly concerning given that recognizing the significance of a medical diagnosis represents a core empathetic skill. The models' failure to adequately identify key events suggests they may systematically underestimate the emotional weight of diagnostic moments for patients. This deficit could contribute to the demographic biases observed in our main analysis, as models that fail to recognize emotional significance may default to stereotypical assumptions about different patient groups' emotional needs.

In contrast, *Intention Recognition* and *Mixed Event Recognition* achieved moderate scores ($\approx$ 2.0), suggesting adequate understanding of explicit communicative intentions. However, this pattern–competent explicit recognition paired with poor implicit recognition–mirrors our finding that cognitive empathy remains stable while affective empathy varies dramatically. The models appear capable of processing explicit information but struggle with the emotional subtleties that distinguish truly empathetic communication.

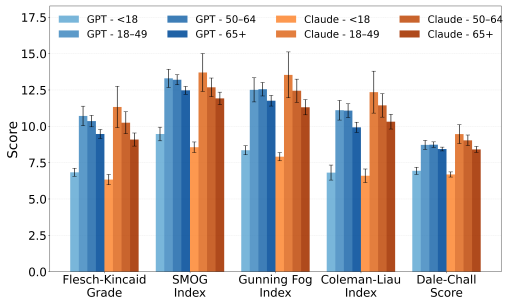
Methodological concerns arise from our discovery of significant self-evaluation bias (GPT consistently inflating its own empathy ratings by 0.333 points), as the use of GPT-4o judges for EmotionQueen evaluation introduces potential systematic bias. The consensus approach may mitigate but not eliminate this concern, particularly since all five judges share the same underlying model architecture and training data.

The profound deficits in implicit emotion recognition have direct implications for medical AI deployment. Patients experiencing serious diagnoses often communicate distress through subtle cues rather than explicit statements. Models that score 1.1/3.0 on implicit emotion recognition may systematically miss opportunities for empathetic response, potentially exacerbating the demographic biases we identified where certain patient groups (e.g., highly educated patients, those with cardiovascular conditions) already receive reduced empathetic communication.

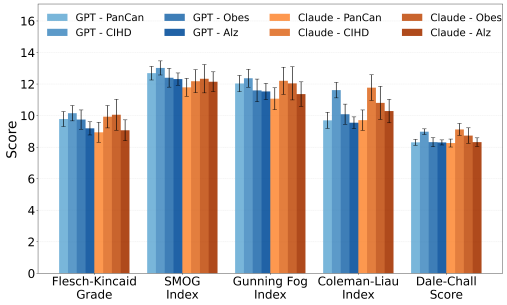
While Claude demonstrated marginally better performance than GPT across most metrics, the differences were minimal (typically $< 0.1$ points) and likely within measurement error. This finding contrasts with our main study's detection of substantial between-model differences in self-evaluation bias, suggesting that EmotionQueen metrics may be less sensitive to the systematic biases we identified through demographic analysis.

The EmotionQueen results thus complement our primary findings by identifying specific empathetic deficits that may underlie the demographic bias patterns observed in our comprehensive analysis. The combination of poor implicit emotion recognition and our documented systematic biases creates compounding risks for equitable empathetic communication across diverse patient populations.

## C Additional understandability figures




**(a)** Understandability scores by age group.




**(b)** Understandability scores by medical condition.

**Figure 7:** Additional analyses: understandability by age group and medical condition. *Condition abbreviations:* Pancreatic cancer (PanCan), Chronic ischemic heart disease (CIHD), Obesity (Obes), Alzheimer's disease (Alz). Bars show Means; error bars denote $\pm$ 95% CIs.

## D Data Overview

| Dataset | Count |
|---|---|
| Claude responses | 156 |
| GPT responses | 156 |
| Total unique responses | 312 |

**Table 2:** Summary of dataset composition

*Note:* Shared prompt numbers: 156 (same demographic scenarios)

## E Score Overview

| Metric | Model | Range | Mean | n |
|---|---|---|---|---|
| Affective Empathy Score | GPT | 1–3 | 2.51 | 312 |
| Affective Empathy Score | Claude | 1–3 | 2.21 | 312 |
| Cognitive Empathy Score | GPT | 2–3 | 2.85 | 312 |
| Cognitive Empathy Score | Claude | 1–3 | 2.81 | 312 |

**Table 3:** Empathy score summary for GPT and Claude

# F Rater Agreement Analysis

## F.1 Claude Responses

| Measure | Correlation | Bias | $p_{\text{BH}}$ | Significant |
|---|---|---|---|---|
| Affective Empathy | $r = 0.46$ | $-0.26$ | $< 0.001$ | Yes |
| Cognitive Empathy | $r = 0.41$ | $+0.03$ | $0.520$ | No |

**Table 4:** Rater agreement for Claude responses

*Note:* Bias calculated as GPT rating minus Claude rating. All $p$-values are Benjamini–Hochberg corrected for multiple comparisons.

## F.2 GPT Responses

| Measure | Correlation | Bias | $p_{\text{BH}}$ | Significant |
|---|---|---|---|---|
| Affective Empathy | $r = 0.28$ | $-0.34$ | $< 0.001$ | Yes |
| Cognitive Empathy | $r = -0.01$ | $-0.11$ | $0.035$ | Yes |

**Table 5:** Rater agreement for GPT responses

*Note:* Bias calculated as GPT rating minus Claude rating. Negative bias indicates GPT rater assigns higher scores than Claude rater. All $p$-values are Benjamini–Hochberg corrected for multiple comparisons.

# G Demographic Bias Analysis

## G.1 Claude Responses Rated by GPT

| Factor | Measure | Statistic | $p_{\text{BH}}$ | Significant |
|---|---|---|---|---|
| Gender | Affective Empathy | $\Delta = +0.05$ ($d = 0.10$) | $0.640$ | No |
| | Cognitive Empathy | $\Delta = +0.10$ ($d = 0.28$) | $0.160$ | No |
| Age | ANOVA | $F(3, 152) = 6.65$ | $0.001$ | Yes |
| | U-shape test | $\Delta = +0.26$ | $0.004$ | Yes |
| Geography | ANOVA | $F = 0.99$ | $0.560$ | No |
| Education | High school vs. Medical | $\Delta = +0.50$ | $< 0.001$ | Yes |
| Diagnosis | ANOVA | $F = 14.28$ | $< 0.001$ | Yes |

**Table 6:** Demographic bias analysis for Claude responses rated by GPT

*Note:* Diagnosis ranking (affective empathy): Alzheimer's highest (2.61), heart disease lowest (1.94). All $p$-values are Benjamini–Hochberg corrected.

### G.2 Claude Responses Rated by Claude

| Factor | Measure | Statistic | $p_{\mathbf{BH}}$ | Significant |
|---|---|---|---|---|
| Gender | Affective Empathy | $\Delta = +0.04$ | 0.770 | No |
| | Cognitive Empathy | $\Delta = -0.01$ | 0.880 | No |
| Age | ANOVA | $F = 1.05$ | 0.560 | No |
| | U-shape test | $\Delta = +0.03$ | 0.840 | No |
| Geography | ANOVA | $F = 0.67$ | 0.640 | No |
| Education | High school vs. Medical | $\Delta = +0.46$ | $< 0.001$ | Yes |
| Diagnosis | ANOVA | $F = 14.06$ | $< 0.001$ | Yes |

**Table 7:** Demographic bias analysis for Claude responses rated by Claude

*Note:* Diagnosis ranking (affective empathy): Pancreatic cancer highest (2.31), heart disease lowest (1.64). All $p$-values are Benjamini–Hochberg corrected.

### G.3 GPT Responses Rated by GPT

| Factor | Measure | Statistic | $p_{\mathbf{BH}}$ | Significant |
|---|---|---|---|---|
| Gender | Affective Empathy | $\Delta = 0.00$ | 1.000 | No |
| | Cognitive Empathy | $\Delta = +0.06$ | 0.410 | No |
| Age | ANOVA | $F = 5.19$ | 0.005 | Yes |
| | U-shape test | $\Delta = +0.26$ | 0.002 | Yes |
| Geography | ANOVA | $F = 0.38$ | 0.780 | No |
| Education | High school vs. Medical | $\Delta = +0.30$ | 0.003 | Yes |
| Diagnosis | ANOVA | $F = 21.81$ | $< 0.001$ | Yes |

**Table 8:** Demographic bias analysis for GPT responses rated by GPT

*Note:* Diagnosis ranking (affective empathy): Alzheimer's highest (2.97), heart disease lowest (2.28). All $p$-values are Benjamini–Hochberg corrected.

### G.4 GPT Responses Rated by Claude

| Factor | Measure | Statistic | $p_{\text{BH}}$ | Significant |
|---|---|---|---|---|
| Gender | Affective Empathy | $\Delta = +0.01$ | 0.910 | No |
| | Cognitive Empathy | $\Delta = -0.05$ | 0.600 | No |
| Age | ANOVA | $F = 1.91$ | 0.230 | No |
| | U-shape test | $\Delta = +0.17$ | 0.110 | No |
| Geography | ANOVA | $F = 2.73$ | 0.140 | No |
| Education | High school vs. Medical | $\Delta = +0.08$ | 0.600 | No |
| Diagnosis | ANOVA | $F = 3.57$ | 0.035 | Yes |

**Table 9:** Demographic bias analysis for GPT responses rated by Claude

*Note:* Diagnosis ranking (affective empathy): Alzheimer's highest, obesity lowest. All $p$-values are Benjamini–Hochberg corrected.

## H Response Source Comparison

| Measure | Claude | GPT | Difference | $p_{\text{BH}}$ |
|---|---|---|---|---|
| Affective Empathy | 2.35 | 2.68 | $-0.33$ | $< 0.001$ |
| Cognitive Empathy | 2.83 | 2.87 | $-0.03$ | 0.600 |

**Table 10:** Comparison of empathy ratings for Claude vs. GPT responses (rated by GPT)

*Note:* GPT responses rated significantly higher on affective empathy only. Difference calculated as Claude minus GPT.

## I Summary of Significant Findings

| Finding | Status |
|---|---|
| Age U-shaped pattern (Claude $\rightarrow$ GPT) | Significant ($p_{\text{BH}} = 0.001$) |
| Medical condition hierarchy | Significant ($p_{\text{BH}} < 0.001$) |
| Education inverse relationship | Significant ($p_{\text{BH}} < 0.001$) |
| Gender bias | Not significant ($p_{\text{BH}} > 0.05$) |
| Geography bias | Not significant ($p_{\text{BH}} > 0.05$) |
| Source effect (affective empathy) | Significant ($p_{\text{BH}} < 0.001$) |
| Source effect (cognitive empathy) | Not significant ($p_{\text{BH}} = 0.600$) |

**Table 11:** Summary of Benjamini–Hochberg corrected results

# J Gender Bias Analysis

| Source | Score Type | Female | Male | Bias | $t$ | $p$ | Sig. |
|---|---|---|---|---|---|---|---|
| Claude + Claude | Affective | 2.13 | 2.04 | $+0.09$ | 1.03 | 0.305 | No |
| Claude + Claude | Cognitive | 2.87 | 2.86 | $+0.01$ | 0.22 | 0.825 | No |
| Claude + GPT | Affective | 2.37 | 2.32 | $+0.05$ | 0.64 | 0.527 | No |
| Claude + GPT | Cognitive | 2.89 | 2.78 | $+0.10$ | 1.72 | 0.087 | No |
| GPT + Claude | Affective | 2.35 | 2.33 | $+0.01$ | 0.15 | 0.882 | No |
| GPT + Claude | Cognitive | 2.74 | 2.77 | $-0.03$ | $-0.37$ | 0.711 | No |
| GPT + GPT | Affective | 2.68 | 2.68 | 0.00 | 0.00 | 1.000 | No |
| GPT + GPT | Cognitive | 2.90 | 2.83 | $+0.06$ | 1.17 | 0.244 | No |

**Table 12:** Independent t-tests for gender bias by model source and score type

*Note:* Bias calculated as Female minus Male. Source notation: Response model + Rater model. No comparisons reached statistical significance ($p < 0.05$).

# K Intra-Model Bias and Evaluation Consistency Analysis

Table 13 shows intra-model self-evaluation patterns, consistency metrics, and demographic-specific bias variance for GPT and Claude, with primary self-evaluation bias observed in affective empathy and minimal bias in cognitive empathy.

| Analysis Category | GPT Pattern | GPT Statistics | Claude Pattern | Claude Statistics | Effect Size | Significance |
|---|---|---|---|---|---|---|
| **Self-Evaluation Bias (Affective Empathy)** | | | | | | |
| Own vs Other Rating | 2.679 vs 2.346 (+0.333 inflation) | $p < 0.0001$ $n = 156$ each | 2.083 vs 2.340 (-0.256 deflation) | $p < 0.0001$ $n = 156$ each | $d = 0.686$ (GPT) $d = -0.473$ (Claude) | ✓ Highly Significant |
| Cognitive Empathy Bias | +0.032 inflation | $p = 0.430$ (n.s.) | +0.109 inflation | $p = 0.016$ (sig.) | $d < 0.3$ (small) | Minimal bias |
| **Within-Model Rating Consistency** | | | | | | |
| Variance Self vs Other | 0.219 vs 0.254 (Ratio: 0.864) | CV: 0.175 vs 0.215 Similar consistency | 0.296 vs 0.290 (Ratio: 1.021) | CV: 0.261 vs 0.230 Similar consistency | Variance ratios within normal range | × No substantial difference |
| **Demographic-Specific Self-Bias Patterns (Affective Empathy)** | | | | | | |
| Gender Groups | Female: +0.308 Male: +0.359 | Both $p < 0.001$ Variance: 0.0007 | Female: -0.244 Male: -0.269 | Both $p < 0.01$ Variance: 0.0002 | Consistent bias across genders | ✓ Significant in all groups |
| geographical group Groups | African: +0.308 Asian: +0.269 European: +0.423 | All $p < 0.01$ Variance: 0.0043 | African: -0.231 Asian: -0.115 European: -0.423 | Af & Eu $p < 0.05$ As $p = 0.288$ Variance: 0.0161 | Strongest bias for European | ✓ Mostly significant |
| Education Groups | HS: +0.217 Univ: +0.396 Med: +0.417 | All $p < 0.05$ Variance: 0.0081 | HS: -0.033 Univ: -0.375 Med: -0.417 | Univ & Med $p < 0.001$ HS $p = 0.732$ Variance: 0.0295 | Bias increases with education | ✓ Significant for higher education |
| Medical Diagnosis | Obesity: +0.208 Alzheimer's: +0.361 Heart: +0.333 Cancer: +0.472 | All $p < 0.05$ Variance: 0.0088 | Obesity: -0.062 Alzheimer's: -0.250 Heart: -0.611 Cancer: -0.167 | Alz & Heart $p < 0.05$ Others n.s. Variance: 0.0426 | Cancer highest bias for GPT | ✓ Varies by condition |
| **Cross-Evaluation and Interaction Effects** | | | | | | |
| Cross-Rater Agreement | GPT rating Claude: 2.346 (aff. emp.) | | Claude rating GPT: 2.340 (aff. emp.) | | $r = -0.032$ Poor agreement on affective empathy | × No asymmetry $p = 0.914$ |
| Response-Rater Matrix | GPT→GPT: 2.679 GPT→Claude: 2.340 (affective empathy) | | Claude→Claude: 2.083 Claude→GPT: 2.346 (affective empathy) | | Largest gap: Claude self vs GPT self ($d = -1.174$) | ✓ Significant interaction effects |

**Table 13:** Comprehensive intra-model bias analysis showing self-evaluation patterns, consistency metrics, and demographic-specific bias variance across GPT and Claude evaluators. Primary self-evaluation bias findings are for affective empathy; cognitive empathy shows minimal bias patterns. HS = High School, Univ = University, Med = Medical degree, Alz = Alzheimer's, Heart = Chronic Ischemic Heart Disease, n.s. = not significant, sig. = significant.

# L Human Evaluation

Table 14 presents paired t-tests comparing human mean scores versus model-generated affective and cognitive empathy scores across demographic categories.

Table 15 reports t-tests comparing affective and cognitive empathy scores between African female and European female groups with high school or lower education.

Table 16 shows a detailed distribution of human ratings, including 95% CI and rating ranges.

| Category | H vs GPT (Aff) | H vs Claude (Aff) | H vs GPT (Cog) | H vs Claude (Cog) |
|---|---|---|---|---|
| European female | $t = -5.25\,\mathbf{p = 0.001}$ | $t = -0.23\,p = 0.820$ | $t = -3.75\,\mathbf{p = 0.005}$ | $t = -1.15\,p = 0.280$ |
| African female | $t = -2.59\,\mathbf{p = 0.029}$ | $t = 0.00\,p = 1.000$ | $t = -3.71\,\mathbf{p = 0.005}$ | $t = -1.29\,p = 0.229$ |
| European male | $t = -3.00\,\mathbf{p = 0.015}$ | $t = -0.69\,p = 0.509$ | $t = -2.45\,\mathbf{p = 0.037}$ | $t = -1.50\,p = 0.168$ |
| African male | $t = -6.00\,\mathbf{p < 0.001}$ | $t = -0.56\,p = 0.591$ | $t = -3.00\,\mathbf{p = 0.015}$ | $t = -1.86\,p = 0.096$ |

**Table 14:** Paired t-test of human mean vs. model empathy (Affective empathy and Cognitive empathy) scores on GPT-generated responses, stratified by demographic category. Bold indicates p < 0.05.

| Source | t-statistic | p-value |
|---|---|---|
| Affective Human | -2.38 | **0.029*** |
| Affective GPT | -1.57 | 0.138 |
| Affective Claude | -1.90 | 0.075 |
| Cognitive Human | -1.56 | 0.135 |
| Cognitive GPT | -1.00 | 0.343 |
| Cognitive Claude | -0.95 | 0.356 |

**Table 15:** T-tests between African female and European female groups (with high school or lower education). * p < 0.05

| Rater | African Female | African Male | European Female | European Male | Affective Mean $\pm$ 95% CI | Affective Range | Cognitive Mean $\pm$ 95% CI | Cognitive Range |
|---|---|---|---|---|---|---|---|---|
| Human 1 | 10 | 10 | 10 | 0 | $2.27 \pm 0.22$ | 1–3 | $2.70 \pm 0.20$ | 1–3 |
| Human 2 | 10 | 0 | 10 | 10 | $2.10 \pm 0.23$ | 1–3 | $2.27 \pm 0.29$ | 1–3 |
| Human 3 | 10 | 0 | 10 | 0 | $2.35 \pm 0.27$ | 1–3 | $2.65 \pm 0.23$ | 2–3 |
| Human 4 | 10 | 0 | 10 | 0 | $2.75 \pm 0.21$ | 2–3 | $2.25 \pm 0.30$ | 1–3 |
| **Total** | **40** | **10** | **40** | **10** | | | | |

**Table 16:** Detailed distribution of human ratings across demographic groups, including affective and cognitive empathy statistics (mean $\pm$ 95% CI, based on sample standard deviation).

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions, which include: (1) developing an evaluation framework for assessing LLM understandability and empathy in medical diagnoses, (2) evaluating GPT-4o and Claude-3.7, (3) identifying systematic biases in both dimensions, and (4) revealing self-evaluation biases in LLM-as-a-judge approaches. The scope is clearly defined as focusing on diagnostic communication rather than clinical accuracy.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses several limitations in the Conclusion section, including: narrow range of patient scenarios, small human evaluation sample size, evaluator bias issues with weak inter-rater agreement, and cultural contingency of empathy judgments. The authors also acknowledge that scenarios cannot capture full real-world variability.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results requiring formal proofs. It is an empirical evaluation study using established readability metrics and empathy assessment frameworks.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides detailed methodology including: prompt templates (Section 3), specific readability metrics with formulas (Table 1), empathy evaluation prompts (code listings), demographic combinations (156 prompts), and statistical analysis methods. The framework and evaluation pipeline are thoroughly documented.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: uploaded in the supplementary material

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports appropriate statistical measures including error bars (±1 SD), p-values, effect sizes (Cohen's d), ANOVA results, and t-test statistics. Comprehensive statistical validation is provided in the appendix with detailed significance testing results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports appropriate statistical measures including error bars (±1 SD), p-values, effect sizes (Cohen's d), ANOVA results, and t-test statistics. Comprehensive statistical validation is provided in the appendix with detailed significance testing results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper mentioned the usage of API calls to the corresponding LLM models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We ensure that our experiments comply with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes a dedicated “Broader Impacts" section discussing both positive impacts (improving healthcare access, supporting clinicians) and negative impacts (amplifying health inequities, undermining trust among vulnerable groups). It emphasizes the need for transparent, bias-aware evaluation before deployment

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate defakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The evaluation framework itself poses no risk and could help implement safeguards in medical AI systems.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses established commercial LLM APIs (properly cited), and the EmotionQueen framework, whose repository is provided in Appendix A as a footnote, and all assets are used in accordance with their terms of use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The paper includes human evaluation with four annotators from the research team.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: no external human participant, only the authors have participated in this project.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs (GPT-4o and Claude-3.7) are the core focus of this research - both as the systems being evaluated and as judges in the LLM-as-a-judge evaluation framework. Their usage is thoroughly documented and central to the methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.