Cross-lingual Evaluation of Diagnostic Accuracy Across Patient sex in LLMs

Hajer Janabi, Anum Khan, Chaturya Pasupuleti, Siddhi Pareek, Raj Davuluri, Ayush Davuluri, Rishit Lalchandani and Sai Prabhav Vankamamidi
DiscoverSTEM, Plano, TX, USA

janabi.hajer@gmail.com
ak3430253@gmail.com
ryapas12@gmail.com
siddhipareek24@gmail.com
rajv.davuluri@gmail.com
ayush.davuluri@gmail.com
rishilalchandani14@gmail.com
spvankamamidi@gmail.com

Abstract: Large language models (LLMs) are increasingly used to answer health questions across languages, making equitable diagnostic performance an important safety concern. This study examined whether patient sex and input language jointly affected diagnostic accuracy in GPT-4o mini, Claude Sonnet 4.6, and Gemini Flash-Lite 3.1. A total of 669 synthetic DDXPlus profiles (340 female and 329 male) were evaluated in English, Spanish, French, Hindi, Arabic, and Mandarin using an automated Python workflow. The primary outcome was exact-match diagnostic accuracy; technical response success and model-reported confidence were secondary outcomes. Across the five non-English languages, Claude achieved the highest diagnostic accuracy (29.9%), followed by Gemini (27.1%) and GPT-4o mini (22.7%). Contrary to the hypothesis, female profiles achieved higher accuracy than male profiles across all models: 32.6% versus 27.2% for Claude, 30.0% versus 24.1% for Gemini, and 26.6% versus 18.8% for GPT-4o mini. Accuracy also varied substantially across languages. These findings did not support the predicted intersectional disadvantage for female profiles, but they demonstrate that multilingual medical LLM performance depends on both the model and evaluation context. Rigorous cross-lingual and demographic validation is needed before these systems are used for clinical decision support.

Keywords: Large language models, Diagnostic accuracy, Patient sex, Multilingual evaluation, AI in healthcare, Intersectional Bias

1. Introduction

Large language models are becoming common tools for clinicians and patients. In 2026, 81% of physicians reported awareness or use of AI, including 18% who used AI translation tools and 17% who used assistive diagnostic tools (American Medical Association, 2026). Health-related inquiries also represent approximately 5% of ChatGPT messages worldwide (OpenAI, 2026). As these systems influence how symptoms and diagnoses are interpreted, their reliability across demographic groups and languages requires careful evaluation. LLMs learn from large datasets that may reproduce existing healthcare disparities (Omar et al, 2025; Omiye et al, 2023). This is particularly relevant to patient sex because women have historically been underrepresented in clinical research, leaving portions of medical evidence and diagnostic criteria disproportionately based on male populations (Warren, Garrett, and Frame, 2025).

Previous research has examined language-related or sex-related differences in medical LLM performance, including evidence that GPT-4 can perpetuate racial and sex biases in healthcare (Zack et al., 2024), but fewer studies have evaluated how these variables interact (Omar et al., 2025; Zakir Sulaiman et al., 2024). Single-factor evaluations may therefore underestimate bias in multilingual clinical settings. This study addresses that gap by comparing diagnostic accuracy across patient sex, language, and model under standardised conditions. We hypothesised that translating otherwise equivalent female profiles would reduce diagnostic accuracy more than translating male profiles, indicating an interaction between sex and language.

2. Methodology

A retrospective experimental design was used to evaluate diagnostic performance across patient sex and language. Three accessible LLMs were tested: GPT-4o mini, Claude Sonnet 4.6, and Gemini Flash-Lite 3.1. These models were selected because they were widely accessible, developed by different organisations, and supported multilingual prompting.

The experiment compared four conditions: male and female profiles presented in English and male and female profiles presented in a translated language. Sex assignment was represented through pronouns, while the remaining case information was held constant within matched profiles. English served as the control condition because it was the primary training language of the evaluated models. The five comparison languages were Spanish, French, Hindi, Arabic, and Mandarin. English prompts were translated using Google Translate's neural machine translation system (Wu et al, 2016), selected for its broad public availability and relevance to real-world use.

The study used 669 synthetic patient profiles from DDXPlus (Fansi Tchango et al, 2022), including 340 female and 329 male profiles. Each model evaluated the profiles across the six language conditions. The sample included acute conditions, such as NSTEMI, unstable angina, pulmonary embolism, and anaphylaxis, as well as chronic or non-emergent conditions, such as anaemia, gastro-oesophageal reflux disease, myasthenia gravis, and sarcoidosis. Only qualitative symptoms, categorical risk factors, and structured diagnostic evidence were retained because these features resemble natural patient descriptions; quantitative vital signs and measurements were excluded. Age and associated clinical features were held constant or randomised across matched conditions where appropriate.

Data collection was automated using a custom Python 3.13.14 script and the official SDKs from OpenAI, Anthropic, and Google. DDXPlus cases were converted to standardised CSV files, and symptom text was translated from English into Spanish, French, Hindi, Arabic, and Mandarin using the GOOGLETRANSLATE function in Google Sheets. Patient identifiers, sex labels, and language tags were preserved. For each case, the script populated one standardised prompt and recorded the model's predicted diagnosis, reported confidence, patient identifier, language, and sex in a master CSV file. ChatGPT assisted only with spreadsheet formulas and formatting; it was not used to translate symptoms, score diagnoses, or conduct the analysis.

The primary outcome was diagnostic accuracy, defined as exact agreement between each model's primary diagnosis and the DDXPlus ground-truth diagnosis. Predictions were coded as correct or incorrect and summarised by model, language, patient sex, and trial. Confidence distributions were retained as a descriptive secondary outcome. A separate technical response-success outcome captured whether a response could be processed for analysis; failures were analysed using Firth-penalised logistic regression with English as the reference language. Missing or unusable outputs were excluded rather than imputed. Results were summarised using counts, percentages, and averages, and visualisations were produced in Google Sheets and Canva (version 4.218.0).

3. Results

The analytic sample comprised 669 DDXPlus profiles evaluated across three LLMs and six languages. The study reports 12,042 model-case evaluations, including 6,120 female-profile and 5,922 male-profile evaluations (Table 1).

Table 1: Experimental dataset and evaluation design.

Variable	Value
Total Cases	12,042 (669 per language per LLM)
Models	3: ChatGPT, Claude, Gemini
Languages	6: English (control), Spanish, French, Hindi, Arabic, Mandarin
Total Male Cases	5,922 (329 per language per LLM)
Total Female Cases	6,120 (340 per language per LLM)
Amount of Trials	Two trials were reported for translated-language conditions; English was evaluated once as the control condition.

Technical response success was 100% for Claude (5,309/5,309) and GPT-4o mini (5,348/5,348), leaving no outcome variance for model-specific regression. Gemini recorded 24 failures out of 8,007 responses (0.3%). In the Gemini analysis, Chinese (OR = 0.13, p = .048), French (OR = 0.13, p = .049), and Spanish (OR = 0.07, p = .004) had lower odds of failure than English. All 24 Gemini failures occurred in the English condition, including 11 female-profile and 13 male-profile evaluations.

Across the five non-English languages, diagnostic accuracy was highest for Claude (29.9%), followed by Gemini (27.1%) and GPT-4o mini (22.7%). Accuracy varied by language and model (Figure 1). Female profiles achieved

higher aggregate accuracy than male profiles across all models: 32.6% versus 27.2% for Claude, 30.0% versus 24.1% for Gemini, and 26.6% versus 18.8% for GPT-4o mini. At the language-by-sex level, the highest accuracies were observed for Claude's Hindi female profiles (41.9%) and Mandarin female profiles (40.9%); the lowest were observed for Gemini's French male profiles (9.9%) and GPT-4o mini's Spanish male profiles (12.7%) (Figure 2).

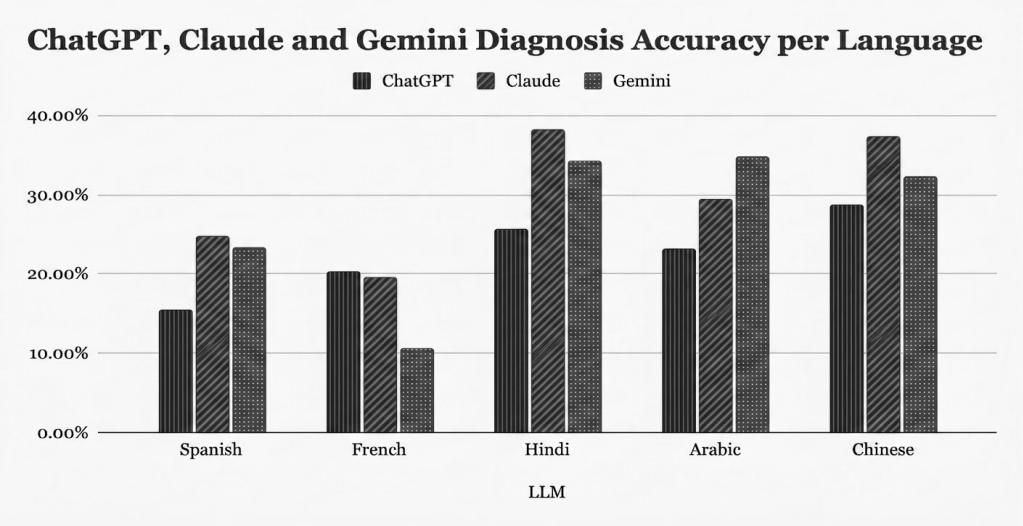

Figure 1: Diagnostic accuracy by model and language. Accuracy was generally higher in Hindi, Arabic, and Mandarin and lower in Spanish and French

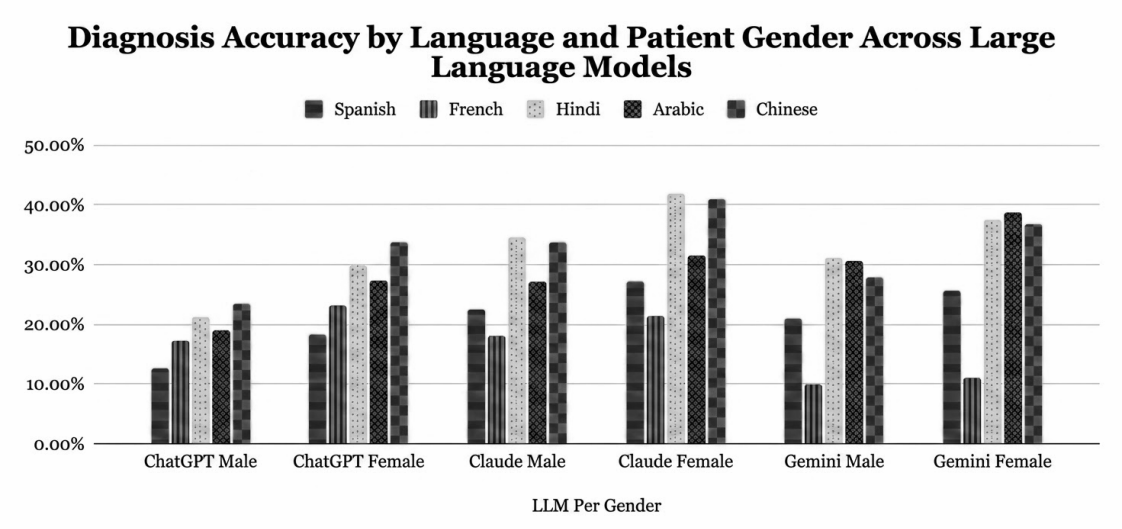

Figure 2: Diagnostic accuracy by model, language, and patient sex. Across the five non-English languages, female profiles achieved higher accuracy than male profiles in each evaluated model

4. Discussion

Contrary to the original hypothesis, female profiles did not experience a larger cross-lingual reduction in diagnostic accuracy than male profiles. Female profiles achieved higher accuracy across all three models, whereas performance varied more visibly across languages and model families. Claude achieved the highest aggregate accuracy, followed by Gemini and GPT-4o mini. However, the consistently higher performance for female profiles should not be interpreted as evidence that these systems favour female patients. Because the profiles were drawn from DDXPlus, differences in disease distribution, symptom specificity, or case difficulty may have contributed to the observed pattern.

The language-level differences reinforce prior findings that medical LLM performance is not uniform across languages. Hindi, Arabic, and Mandarin generally produced higher diagnostic accuracy than Spanish and French, although the pattern differed by model. Exact-match scoring provided a consistent automated comparison, but

it may classify clinically equivalent wording as incorrect. The near-complete technical response success rates also indicate that receiving a usable output is distinct from receiving the correct diagnosis.

Several limitations constrain interpretation. The study evaluated three LLMs using synthetic rather than real patient profiles, only six languages, machine-translated prompts, and a restricted set of conditions. It did not include clinician review of translation quality or partially correct diagnoses. Repeated evaluations of the same profiles may also violate the independence assumed by simple percentage comparisons. Accordingly, the results should be understood as a benchmark evaluation rather than evidence of clinical readiness or a definitive estimate of demographic bias. Future work should use expert-validated translations, real or independently curated clinical cases, clinically informed scoring, and statistical models that directly test sex-by-language interactions while accounting for repeated profiles.

5. Conclusion and Future Work

This study evaluated whether patient sex and input language jointly affected diagnostic accuracy across three widely accessible LLMs. The hypothesised disadvantage for female profiles was not observed: female profiles achieved higher diagnostic accuracy than male profiles across every model, while accuracy varied substantially by language and model. Claude produced the highest overall accuracy, although no model exceeded 30% across the non-English conditions. These findings do not establish that LLMs are free from sex-related bias; instead, they show that benchmark composition, language, and model choice can materially affect measured performance. Before medical LLMs are used for multilingual diagnostic support, they should be validated across representative patient groups and languages. Future studies should use larger clinical datasets, expert-reviewed translations, clinically informed diagnosis matching, and inferential models that directly test sex-by-language interactions.

Acknowledgements

Thank you to DiscoverSTEM, a non-profit organisation that offers college scholarships to young students and innovators, which is the site at which experiments were performed, data was gathered, and the framework for the whole project was provided.

Ethics Declaration: This study used publicly available synthetic patient profiles and did not involve human participants or animals; therefore, informed consent was not applicable. The work was conducted under the oversight of DiscoverSTEM in Plano, Texas, from June to August 2026. Formal institutional review board approval was not required for this evaluation of publicly available LLMs and synthetic data.

AI Declaration: Artificial intelligence tools were used to support language editing, proofreading, and figure development. The authors independently verified all data, analyses, interpretations, citations, and final manuscript decisions.

References

American Medical Association (2026). 2026 Physician Survey on Augmented Intelligence, Center for Digital Health and AI. Available at: https://www.ama-assn.org/system/files/physician-ai-sentiment-report.pdf

Fansi Tchango, A., Goel, R., Wen, Z., Martel, J. and Ghosn, J. (2022) “DDXPlus: A Rich Dataset for Automatic Medical Diagnosis”, Advances in Neural Information Processing Systems, Vol 35, pp 31306–31318.

Omar, M., Sorin, V., Agbareia, R. et al (2025) “Evaluating and Addressing Demographic Disparities in Medical Large Language Models: A Systematic Review”, International Journal for Equity in Health, Vol 24, Article 15.

Omiye, J.A., Lester, J.C., Spichak, S., Rotemberg, V. and Daneshjou, R. (2023) “Large Language Models Propagate Race-Based Medicine”, NPJ Digital Medicine, Vol 6, No. 1, Article 195.

OpenAI (2026) AI as a Healthcare Ally: How Americans Are Navigating the System with ChatGPT, January. Available at: https://cdn.openai.com/pdf/2cb29276-68cd-4ec6-a5f4-c01c5e7a36e9/OpenAI-AI-as-a-Healthcare-Ally-Jan-2026.pdf

Warren, A., Garrett, K. and Frame, L.A. (2025) “Disparities in Women's Health and Clinical Considerations from a Translational Science Perspective: A Narrative Review and Framework for Future Directions”, Women's Health, Vol 21.

Wu, Y., Schuster, M., Chen, Z. et al (2016) “Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”, arXiv.

Zack, T., Lehman, E., Suzgun, M. et al (2024) “Assessing the Potential of GPT-4 to Perpetuate Racial and Sex Biases in Health Care: A Model Evaluation Study”, The Lancet Digital Health, Vol 6, No. 1, pp e12–e22.

Zakir Sulaiman, M. et al (2024) “Evaluating Intersectional Bias and Diagnostic Disparities in Clinical Large Language Models”, Journal of the American Medical Informatics Association, Vol 31, No. 5, pp 1102–1111.