Journal of Asian Health

ORIGINAL RESEARCH ARTICLE

Large Language Models and Asian American Gastric Cancer Risk

Gloria Wu[1], Aadjot Sidhu[2], Sahej Sidhu[3], Hrishi Paliath-Pathiyal[4], Obaid Khan[5], Milan del Buono[6], Peter C. Lo[7]

Introduction: This study assessed the ability of five large language models (LLMs) to convey information about gastric cancer to Asian American patients.

Methods: A series of questions in six languages was posed to five chatbots (ChatGPT-3.5, ChatGPT-4o, Gemini, Claude, and Coral) regarding gastric cancer and its incidence among Asian American subpopulations. Using a two-way analysis of variance, the AI (artificial intelligence) self-rated score, the human evaluator score, and the manual scores per model and language were analyzed to detect any significant differences among the LLMs.

Results: Significant differences in performance were observed with Claude, Gemini, and ChatGPT-4o when compared with Coral across all of the languages tested (p adjusted = 0.008, 0.038, and 0.008, respectively). Among these chatbots, Claude and GPT4o were found to outperform GPT-3.5 (p adjusted = 0.038 and 0.007, respectively). T-tests across all six languages revealed significant differences between Chinese versus Punjabi and English versus Korean, Punjabi, and Vietnamese (p adjusted = 0.042, 0.027, 0.025, and 0.025, respectively).

Conclusion: Additional fine-tuning and the development of more diverse language datasets would allow these LLMs to address information about gastric cancer health disparities among Asian American subpopulations. LLMs may also play a crucial role in bridging education gaps within these communities, enhancing overall health literacy and contributing to better health outcomes.

Key Words: AI ▪ large language models ▪ Asian American ▪ gastric cancer ▪ stomach neoplasms

In the United States, the prevalence of gastric cancer among Asian Americans has been poorly recognized by healthcare professionals,[1] despite noteworthy ethnic differences and gender disparities. Although gastric cancer is considered to be a rare disease in the US, studies indicate that men are twice as likely to develop gastric cancer as women,[2] with Asian American men at high risk for the disease.[3] Gastric cancer is among the top five most diagnosed cancers in the world and is the third leading cause of death among Asian persons worldwide.[4,5] According to the American Cancer Society, Asian Americans, who comprise 7% of the US population, are twice as likely to develop gastric cancer than their White counterparts.[2,6] Furthermore, Korean Americans experience the highest rates of gastric cancer, with over five times the risk of developing gastric cancer compared with White Americans,[4] but are screened at the same rate.

The US government has implemented cost-effective colon cancer screening programs; however, these lack specific guidelines for gastric cancer that might apply to both general populations and high-risk Asian Americans.[7] Moreover, private insurance companies might be less inclined to cover Asian American patients for endoscopic screening and other preventive measures based on low prevalence rates among White populations as the benchmark for their decision-making. Early detection of gastric cancer in Asian

Correspondence to: Gloria Wu, 2550 Samaritan Dr., Suite C, San Jose, CA 95124, USA. Email: gwu2550@gmail.com

Journal of Asian Health is available at https://journalofasianhealth.org

POPULAR SCIENTIFIC SUMMARY

- Asian Americans, particularly men, are at high risk for gastric cancer. According to the American Cancer Society, Asian Americans, who comprise 7% of the United States population, are twice as likely to develop gastric cancer as their White counterparts. Despite being at high risk, Asian Americans are screened at the same rate as other patients. Therefore, gaining language appropriate knowledge about the risks and available screening measures is the key to prevention among Asian Americans.
- Five chatbots were tested using six Asian languages to assess how well each large language model was able to respond to questions about gastric cancer in each language. While two of the chatbots responded clearly, the others provided uneven results. With further fine-tuning and development of more diverse language datasets, these large language models might improve health literacy among Asian subpopulations and lead to improved health outcomes.

Americans through endoscopy would improve survival rates and potentially save lives by preventing the progression of the disease.[8,9]

Given inadequate public awareness about the high incidence of gastric cancer among Asian Americans and limited access to screening measures,[10,11] at-risk individuals might turn to artificial intelligence (AI) chatbots for medical information. Theoretically, these free and accessible chatbots can provide crucial information for those who are experiencing barriers to adequate healthcare and coverage.

However, the paucity of health data about Asian Americans in AI datasets might hinder the accuracy of search results.[12] Recent advancements in machine learning methodologies hold promise for enhancing the performance and accuracy of AI chatbots for complex medical inquiries. Interestingly, the AI industry has seen substantial progress, with chatbots like Open AI's ChatGPT (generative pre-trained transformer) and Google's Gemini continuously refining their databases through millions of user interactions driven by two primary machine learning approaches: unsupervised and supervised learning.

Unsupervised AI learning allows models to identify patterns and structures in large datasets without human supervision. This approach sifts through a wide range of references, searches for key terms, and clusters the repetitive information to form an output.[13] However, the lack of human guidance and inherent nationality bias can lead to misinterpretations, which, in turn, can lead to errors of omissions or inaccuracies about gastric cancer among Asian Americans.[14]

The aim of this study was to evaluate the ability of five different AI-driven large language model (LLM) chatbots to deliver accurate and coherent information on gastric cancer in non-English languages. Multicultural patients and their families in the US often rely on AI for medical information because of cultural, linguistic, and financial barriers.[13] The present study sought to determine whether chatbots can be a reliable source of crucial medical information for Asian Americans with gastric cancer.

METHODS

This study evaluated five chatbots (ChatGPT-3.5 [Open AI], ChatGPT-4o [Open AI], Gemini [Google], Claude [Anthropic], and Coral [Google]) by posing questions about gastric cancer in six languages: English, Chinese, Vietnamese, Korean, Punjabi, and Hindi (Supplemental Table 1). The questions included: (1) 'What is gastric cancer?'; (2) 'I am a 40-year-old male, experiencing bloating and loss of appetite'; (3) 'Who is at risk of gastric cancer in America?' The six languages chosen varied in syntax and alphabets. The English, Chinese, Vietnamese, Korean, Hindi, and Punjabi languages possess different phonology, grammar structure, and script. Although the Vietnamese language uses Latin script, it was included for its extensive use of complex diacritical marks and tones. Additionally, Punjabi and Vietnamese, languages with fewer native speakers, were included to assess whether languages with more speakers performed better with the chatbots. This study excluded the Japanese language because Japanese Kanji characters have the same syntax as the Chinese language and uses very similar characters.

Table 1. Chatbot Self-Assessment and Human Evaluation of Query Results by Readability, Fluency, and Accuracy

Criteria	Score 1	Score 2	Score 3	Score 4	Score 5
Readability	Extremely difficult to read, poor grammar/ syntax	Difficult to read, multiple grammar errors	Moderately readable, some grammatical issues	Easy to read, minor grammatical errors	Excellent readability, perfect grammar
Fluency	Extremely choppy, unnatural language flow	Somewhat choppy, noticeable language issues	Moderately fluent, some unnatural phrasing	Good fluency, minor issues with natural flow	Excellent fluency, sounds like natural native speech
Accuracy	Completely inaccurate medical information	Mostly inaccurate with some correct elements	Partially accurate, missing key information	Mostly accurate with minor omissions	Completely accurate and comprehensive

Both chatbot self-assessment and human evaluation of the query results used identical criteria, rating responses on readability, fluency, and accuracy using a scale of 1–5 (Table 1).

Each chatbot generated three discrete scores for its own response, which were combined using a weighted formula to create a composite score: total score = 0.2 × readability score + 0.4 × fluency score + 0.4 × accuracy score. Two native speakers per language independently evaluated the same responses using the identical three criteria and scoring scale, with human evaluators (GW, AS, SS, HP, OK, MDB, and PCL) blinded to the chatbot's self-ratings during the scoring process to prevent bias. Native speaker evaluations were validated through a dual-reviewer process in which two native speakers per language independently assessed and cross-checked the evaluations, with additional oversight provided by a team of clinicians, surgeons, and ophthalmologists.

The differences in chatbot performance were analyzed using two individual two-way analysis of variance (ANOVA). One ANOVA was performed on the difference between the AI self-rated total score and the human evaluator total score by the model used and the language queried. This analysis determined whether there were differences in how accurately each chatbot rated itself. The second ANOVA was performed on the manual scores alone versus the model used and language queried, to determine whether either parameter played a role in the quality of the chatbot's answers.

Following significant ANOVA results, paired *t*-tests were used as a post-hoc test. The *p*-values from these tests were adjusted using the Benjamini-Hochberg procedure: *p* adjusted = $p \times m/r$, where m is the total number of tests and r is the relative rank of the *p*-value. The Benjamini-Hochberg procedure is advantageous because it controls the false discovery rate without excessively reducing the test's statistical power.

This study was determined to be exempt from institutional review board approval as it involved evaluation of publicly available AI chatbot responses without collection of human subject data. No identifiable personal information was collected during the study.

RESULTS

Meaningful responses in English and Chinese tended to be longer than responses in Hindi, Korean, Punjabi, and Vietnamese. The long responses in Punjabi and Vietnamese contained lines of meaningless text. Coral's responses in Punjabi contained frequent repetitions of the word 'hunting'. Additionally, several chatbots provided fewer than two sentences and were too brief to be informative. Specifically, Coral in Korean and ChatGPT-3.5 in Punjabi had shorter responses, with Coral in Korean having a 44-word count and ChatGPT-3.5 in Punjabi providing one sentence with 26 words. Gemini and ChatGPT4o tended to create responses with longer word counts than Claude, Coral, and ChatGPT3.5 across all languages (Figure 1).

A two-way ANOVA was run on the difference between the AI score and the human score versus the model queried, the language used, and the model queried combined with the language used (Tables 2 and 3).

The ANOVA returns for the model, language, and model + language groups were statistically significant (F = 6.114, 1.215, and 1.743 and p = 1.72e–7, 0.0402, and 6.62e–5, respectively). *T*-tests run with the Benjamini-Hochberg correction reveal that Claude, GPT4o, and Gemini outperformed Coral across all

Figure 1. Average Word Count by Language Queried and Chatbot Model

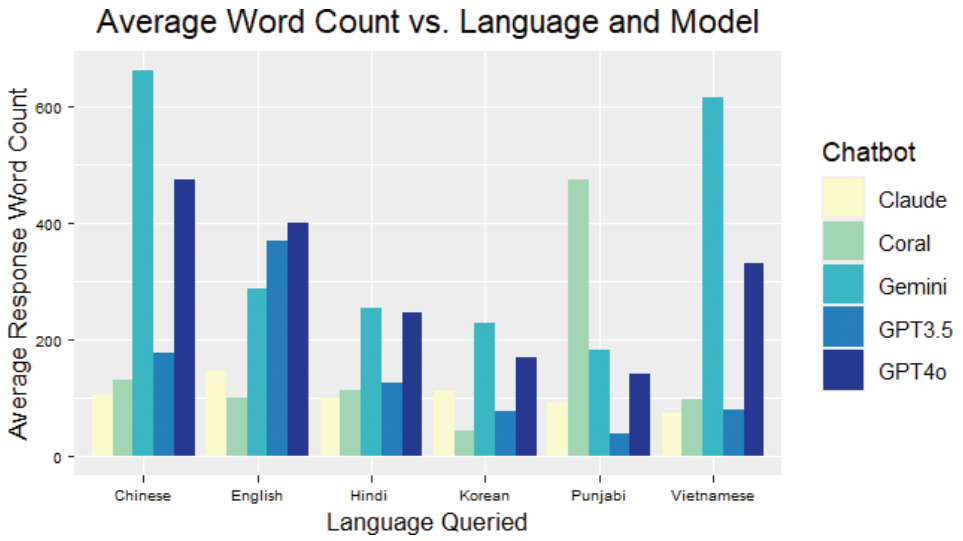


Table 2. Human-rated Scores by Model Queried and Language Used

	ChatGPT-3.5			ChatGPT-4o			Gemini			Claude			Coral		
Human rated	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
English	4.8	4.8	4.6	5	4.8	4.4	5	4.8	5	4.4	4.6	4.8	4.2	3.6	4.6
Chinese	3	3.6	5	5	5	5	5	4.2	4	4	5	5	1.8	3	1.8
Vietnamese	4.6	3.6	4.2	5	4.2	5	2.4	4.4	4	4.6	3.8	4.2	2.4	2.6	3.6
Korean	4.4	4	3.6	4.2	4	4.8	3.8	4	4.6	5	4.4	5	3	3.6	3.6
Punjabi	2.2	1.6	1	4	3.6	4	5	5	3.6	4.6	5	5	1	1	0
Hindi	4	4.8	4	5	4	5	4	4	4	4	3.8	4.6	4.2	4	5

Table 3. AI Self-rated by Model Queried and Language Used

	ChatGPT-3.5			ChatGPT-4o			Gemini			Claude			Coral		
AI self-rated	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
English	5	4.4	4.8	4.4	4.8	4.8	4.4	5	4.6	4	4.6	5	5	5	5
Chinese	4.2	4.4	4.8	4.4	4.8	4.8	4.8	4.4	4.4	4.4	4.4	4.4	5	5	5
Vietnamese	4.2	5	4.4	4.4	4.4	5	5	5	5	4	4.4	4.4	5	5	5
Korean	4.2	4.8	4.4	4.4	4.4	4.4	5	4.4	4.4	5	5	4.8	5	5	5
Punjabi	4.8	4.4	4.4	4.4	4	4.4	4	4.2	4.4	4.4	4.4	4	1	5	0
Hindi	3.4	4.4	5	5	5	4.6	4.6	4.6	4.6	4.4	4.6	4.8	4	4	4

Abbreviations: AI = artificial intelligence.

languages tested (p adjusted = 0.008, 0.038, and 0.008, respectively) (Figure 2). Corrected T-tests did not identify any particular language that had a greater score difference than the other, across chatbots.

A two-way ANOVA was run on the difference between the AI score and the human score versus the model queried, the language used, and the model queried combined with the language used (Table 4).

The ANOVA returns that the model, language, and model + language groups were statistically significant (F-statistic = 8.296, 3.908, and 1.709; p = 4.06e–4, 1.9e–9, and 8.29e–9, respectively). T-tests run with the Benjamini-Hochberg correction reveal that Claude, Gemini, ChatGPT-3.5, and ChatGPT-4o outperformed Coral across languages (p adjusted = 0.002, 0.004, 0.004, and 0.00042, respectively). Additionally, Claude and ChatGPT-4o outperformed ChatGPT-3.5 (p adjusted = 0.038 and 0.007, respectively). T-tests run between languages, across chatbots used, revealing significant differences between Chinese versus Punjabi, and English versus Korean, Punjabi, and Vietnamese (p adjusted = 0.042, 0.027, 0.025, and 0.025, respectively).

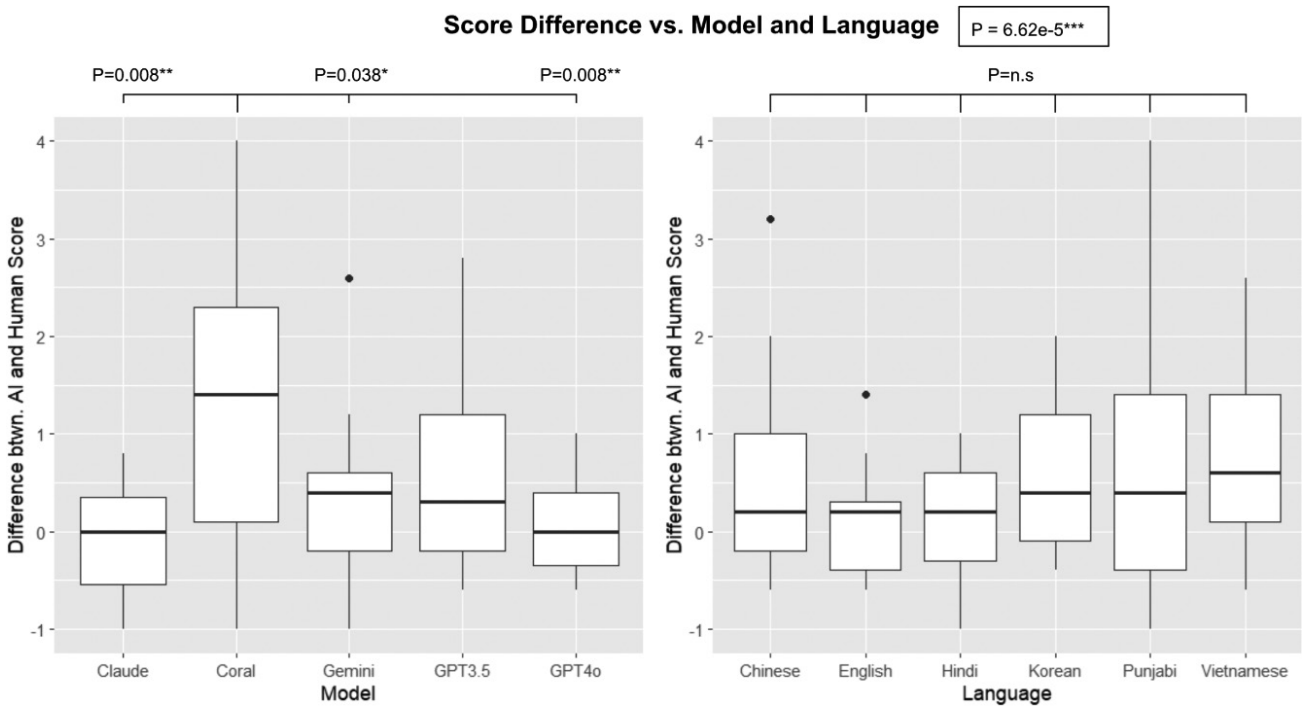
The responses generated by the different AI chatbots differ based on their average word counts. Chatbots Gemini and ChatGPT-4o yield comparatively higher average word counts than ChatGPT-3.5, Gemini, Coral, and Claude (Figure 1). A higher average word count does not correlate to a more thorough response. Coral's Punjabi acts as an outlier that repeats the word 'of hunting' numerous times in its response, increasing the average word count (Figure 2). Consequently, Punjabi is scored the lowest by native speakers, showcasing the inaccuracy of a larger word response (Figure 3). On the other hand, Claude's response, one of the smallest average word counts, yields the second-highest accuracy.

DISCUSSION

This study analyzed the efficacy of five different LLMs in relaying accurate information regarding gastric cancer to Asian Americans who otherwise might not have access to precise medical knowledge. Based on the information available on the internet, this study focused on three main prompts covering the nature of the disease, associated symptoms, and incidence risk among Asian American communities. Six Asian languages were used. This study elected to exclude the Japanese language as its Kanji form uses Chinese characters and because the Japanese American population in the US is among the lowest of all Asian American ethnicities.[14]

Among the five LLMs, significant discrepancies were observed in the chatbots' performance per language tested. While Korean had among the lowest word counts across all six languages, it had one of the highest human reader rating scores, thus proving that there is no correlation between a higher word count and a more accurate response. The same trend occurred with Vietnamese and Chinese languages, with an elevated word count of over 600 words and a high AI vs. human score difference. A similar trend was observed with the five LLMs tested. For Punjabi, the LLM Coral had the highest word count among the other tested chatbots. However, when looking at the AI and human score

Figure 2. Difference between AI Self-rated Scores and Human-rated Scores versus the Model Queried and Language Used[a,b]



Abbreviations: AI = artificial intelligence; n.s. = not significant; IQR = interquartile range.

[a]Outliers were defined as 1.5 × IQR above the third quartile or below the first quartile.

[b]Statistical significance: *$p < 0.05$; **$p < 0.001$; ***$p < 0.001$.

Table 4. Elements of Two-Way ANOVA Comparison of AI Score and Human Score

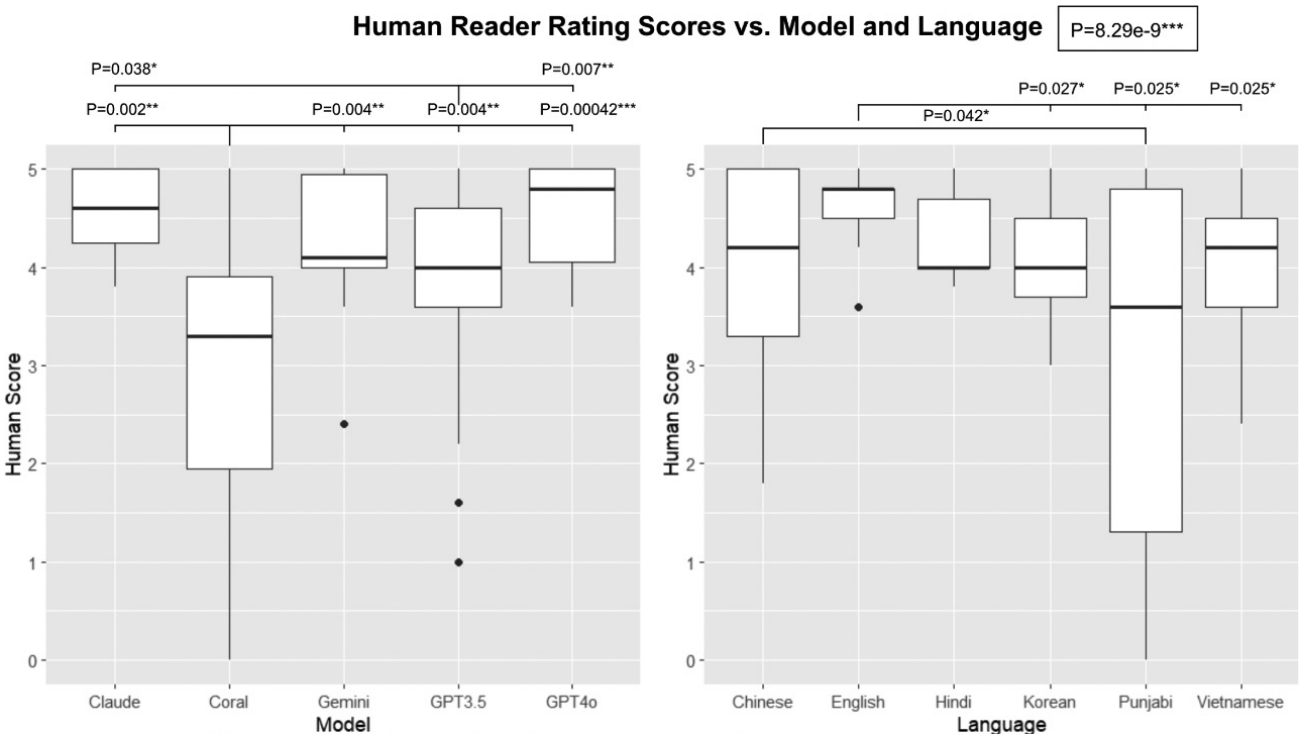
Category	Element compared	Details	Statistical test(s) applied	Purpose
Model	ChatGPT-3.5 vs ChatGPT-4o vs Gemini vs Claude vs Coral	Five LLM chatbots tested for performance	Two-way ANOVA (model factor); post-hoc paired *t*-tests with Benjamini–Hochberg correction	Assess differences in human-rated scores between models
Language	English, Chinese, Vietnamese, Korean, Punjabi, Hindi	Six languages covering diverse scripts and syntactic structures	Two-way ANOVA (language factor); post-hoc paired *t*-tests with Benjamini–Hochberg correction	Assess differences in human-rated scores between languages
Metrics (Human-rated Scores)	Composite score = 0.2 × Readability + 0.4 × Fluency + 0.4 × Accuracy	Readability, Fluency, Accuracy rated on 1–5 scale by native speakers	Compared across models, languages, and model × language	Quantify performance based on native speaker evaluations
Model × Language	All combinations of 5 models × 6 languages	30 unique model-language pairs	Two-way ANOVA interaction term; post-hoc paired *t*-tests with Benjamini–Hochberg correction	Identify interaction effects between model and language
Model × Language × Score	Each unique combination of model, language, and human-rated score	Integrates model performance, language differences, and scoring outcomes	Two-way ANOVA across all factors; post-hoc paired *t*-tests with Benjamini–Hochberg correction	Evaluate how the combination of model and language influences human-rated performance scores

Abbreviations: AI = artificial intelligence; ANOVA = analysis of variance; LLM = large language model.

difference, Coral had the highest degree of variability, hence spotlighting the urgent need for language-specific considerations and larger datasets when querying LLMs for the dissemination of medical information. Among Korean American readers, higher human reader ratings suggest the information was understandable. It is essential to provide clear and concise information about gastric cancer to a population that is five times more susceptible. However, among the LLMs, discrepancies in word count and accuracy across the different languages showcased the underlying health disparity regarding the medical information about gastric cancer that could negatively impact Asian Americans.

The LLM Coral stands out by having a much lower human reader-rated median score compared with the other chatbots (Figure 3). Additionally, Coral demonstrates a large

Figure 3. Human-rated Scores versus the Model Queried and Language Used[a,b]



Abbreviations: IQR = interquartile range.

[a]Outliers were defined as 1.5 × IQR above the third quartile or below the first quartile.

[b]Statistical significance: *$p < 0.05$; **$p < 0.001$; ***$p < 0.001$.

degree of variability and the most significant difference between the AI and human scores, suggesting that human evaluations of Coral's performance are inconsistent, as seen by the wide range of scores. This difference highlights the potential flaws of the AI's assessment capabilities for Coral, compared with the other AI chatbots, which demonstrated more consistent performance metrics.

Additionally, ChatGPT-3.5 exhibited two outliers in the human reader scores (Figure 3). The outliers suggest that, in certain instances, human evaluations of ChatGPT-3.5's performance differ significantly from the majority of the scores. While ChatGPT-3.5 had among the lowest AI and human score differences (Figure 2), it contained two outliers (Figure 3), indicating that ChatGPT-3.5 and similar AI chatbots need to be trained with more linguistically diverse datasets.

However, even with the implementation of larger, diverse datasets, the accuracy of many chatbots may stay the same. Using an unsupervised learning approach to querying various chatbots, this study found that the answer quality tended to be more linguistically accurate. Without imposing specific guidelines on the propagation of medically accurate information, chatbots will resort to continuously generating linguistically accurate responses by focusing on the quality of the language used, readability, and fluency rather than the dissemination of accurate medical information.[15,16]

The fundamental unit of data processed by AI models are 'tokens' that work alongside millions of parameters to generate responses.[17] Tokens can represent characters, words, or even sentences, depending on the language, and each chatbot has a specific maximum token limit per response.[18] This limitation significantly impacts response quality across different languages, as languages with complex writing systems may require more tokens to convey the same information.

Recent advancements have shown remarkable progress in multilingual support and token efficiency. For example, ChatGPT-4o has dramatically improved its handling of non-English languages, reducing the token requirements by 2.9x (from 90 to 31 tokens) for the same content when translating from English to Hindi.[19] This improvement means that a simple greeting like 'Hello, my name is ChatGPT-4o. I'm a new type of language model, it's nice to meet you!' now requires 2.9x fewer tokens when rendered in Hindi as 'नमस्ते, मेरा नाम जीपीटी-4o है। मैं एक नए प्रकार का भाषा मॉडल हूँ। आपसे मलिकर अच्छा लगा!'

This token efficiency improvement has practical implications for medical information delivery. The token context window – the span of tokens a chatbot can process to understand query context and generate responses – becomes more effective when fewer tokens are needed per word. Languages that previously required

more tokens per word produced shorter, less comprehensive responses because all LLM queries have an upper limit on token usage.[20] When token efficiency improves, each token can effectively 'purchase' more words within the context window, allowing for more detailed and informative medical responses in underrepresented languages.

Generative models predict the next 'concept' in a response by calculating the probability distribution of potential tokens, a process often called 'temperature'.[21] This parameter is crucial in balancing the tradeoff between creativity and coherence in model outputs. Higher temperature values cause the model to sample from a flatter probability distribution, considering a broader range of possible tokens. This increases the variability of responses, leading to more diverse outputs. However, this also raises the likelihood of generating less coherent or inaccurate content, as the model might choose less probable concepts not pertinent to the topic.[22] Conversely, lower temperature values steep the probability distribution, narrowing the range of considered tokens and leading to more predictable and uniform responses. This predictability is beneficial in contexts requiring precise and reliable information, like medical data, as it reduces the chances of the model producing outliers or errors. However, these responses lack dynamic, human-like qualities that make interactions feel natural.

Research is needed to find the optimal 'temperature' to avoid artificial hallucinations while maintaining human-like conversation. Developers often focus on balancing these extremes to benefit from both approaches. By fine-tuning the temperature setting, developers aim to achieve a sweet spot in which the responses are creative and coherent.[23] Achieving this balance is not straightforward and requires iterative testing and refinement. By adjusting the temperature and evaluating the outputs, developers can identify the optimal setting for specific applications. This process balances the tradeoff between the desired level of creativity and the acceptable level of coherence. The right temperature allows the model to generate engaging, diverse responses while maintaining accuracy and reliability, enhancing the overall performance and user experience.

While there are readily available medical chatbots such as Google's Med-PaLM 2, they can only be accessed with a paid subscription like Google Cloud, which can pose difficulties for patients who cannot otherwise afford them.[24] Therefore, adding a degree of tunability to publicly available AI will expand its capabilities and ultimately lead to enhanced responses. Since many users lack the knowledge of proper prompt engineering for specific tasks, chatbots often have significant discrepancies in the information they provide. By creating an option to fine-tune the parameters of LLMs trained by a greater, unique, and accurate dataset, the need for prompt engineering from users will be minimized, as will the variability in the data given out by AI. This will allow the public, including Asian Americans who are negatively impacted by the lack of culturally aligned medical information as a significant health disparity, to overcome this barrier without needing extensive training on how to craft effective AI queries for their everyday health questions.

CONCLUSION

This study demonstrated significant variability in LLM performance across different languages when providing gastric cancer information to Asian American populations. While models like Claude, ChatGPT-4o, and Gemini showed superior performance compared with Coral, substantial disparities remain across languages. Underrepresented languages such as Punjabi and Vietnamese showed notably weaker performance. These findings highlight the urgent need for improvements in AI-assisted healthcare communication to address health disparities affecting Asian American communities.

To address token limitations in future chatbot development, several strategic approaches could be implemented. First, dynamic token allocation systems could be developed, which automatically adjust response length based on query complexity and language requirements, allowing more comprehensive answers for complex medical topics. Second, implementing multi-turn conversation capabilities would enable chatbots to provide initial responses within token constraints while offering follow-up interactions for more detailed information. Third, language-specific token optimization could maximize information density for each target language by developing specialized tokenization methods that account for linguistic differences in character-to-meaning ratios. Additionally, hierarchical response structuring could prioritize essential medical information within initial token limits while providing expandable sections for additional details.

Model retraining could significantly enhance responses for underrepresented languages through targeted, culturally informed approaches. Cloud-based platforms like AWS chatbot (Amazon Web Services) provide scalable infrastructure for continuous model improvement through user interaction data. These platforms can collect and analyze user prompts in underrepresented languages, identifying common medical queries and response patterns to inform targeted retraining efforts. Comprehensive data collection efforts should focus on gathering high-quality medical content in underrepresented languages from authoritative sources. This includes partnerships with medical institutions serving Asian American communities to access validated medical translations and culturally appropriate health messaging.

Specialized fine-tuning using domain-specific medical datasets in each target language, combined with reinforcement learning from human feedback (RLHF)

specifically calibrated for medical accuracy, could help models prioritize clinical precision. User prompt analysis can reveal language-specific communication patterns and cultural preferences, enabling more nuanced model adjustments. Implementing iterative evaluation cycles using native speaker medical professionals would ensure that retrained models maintain both linguistic appropriateness and medical accuracy. Additionally, incorporating cultural health beliefs and communication preferences specific to each Asian American subpopulation could improve patient understanding and engagement.

Further adjustment and training with larger, more diverse language datasets would allow these LLMs to address information about gastric cancer health disparities within Asian American subpopulations. Moreover, LLMs may also play a crucial role in bridging education gaps within these communities, enhancing overall health literacy and contributing to better health outcomes through more equitable and culturally competent AI-assisted healthcare communication tools.

ARTICLE INFORMATION

Received October 1, 2024, revised July 30, 2025, accepted August 11, 2025, published January 24, 2026.

Affiliations

[1]Department of Ophthalmology, School of Medicine, University of California, San Francisco, CA, USA; [2]Department of Anthropology, College of Letters and Science, University of California, Davis, CA, USA; [3]Department of Biology, Santa Clara University, Santa Clara, CA, USA; [4]Department of Biological Sciences, Halmos College of Arts & Sciences, Nova Southeastern University, Fort Lauderdale, FL, USA; [5]College of Osteopathic Medicine, California Health Sciences University, Clovis, CA, USA; [6]Department of Engineering, University of California, Berkeley, CA, USA; [7]Department of General Surgery, El Camino Health/Mountain View Surgery, Mountain View, CA, USA

Conflicts of interest and funding sources

The authors have no conflicts of interest or funding sources to disclose.

REFERENCES

1. Taylor VM, Ko LK, Hwang JH, Sin M, Inadomi JM. Gastric cancer in Asian American populations: a neglected health disparity. *Asian Pac J Cancer Prev*. 2014;15(24):10565–71. doi: 10.7314/apjcp.2014.15.24.10565
2. Rawla P, Barsouk A. Epidemiology of gastric cancer: global trends, risk factors and prevention. *Prz Gastroenterol*. 2019;14(1):26–38. doi: 10.5114/pg.2018.80001
3. Shah SC, McKinley M, Gupta S, Peek RM, Martinez ME, Gomez SL. Population-based analysis of differences in gastric cancer incidence among races and ethnicities in individuals age 50 years and older. *Gastroenterology*. 2020;59(5):1705–14.e2. doi: 10.7314/apjcp.2014.15.24.10565
4. American Association for Cancer Research. *AACR Cancer Disparities Progress Report 2022*. 2022. Available from: http://www.CancerDisparitiesProgressReport.org/ [Accessed October 1, 2024].
5. Mukkamalla SKR, Recio-Boiles A, Babiker HM. *Gastric Cancer*. NCBI Bookshelf; 2023. Available from: https://www.ncbi.nlm.nih.gov/books/NBK459142/ [Accessed October 1, 2024].
6. Budiman A, Ruiz NG. *Key Facts about Asians Origin Groups in the U.S.* www.pewresearch.org; 2021. Available from: www.pewresearch.org/short-reads/2021/04/29/key-facts-about-asian-americans/ [Accessed October 1, 2024].
7. Hyun C, Cho D. Gastric cancer disparities in the United States: overcoming the barriers. *Int J Clin Med*. 2024;15(1):19–30. doi: 10.4236/ijcm.2024.151002
8. McMenamin SB, Pourat N, Lee R, Breen N. The importance of health insurance in addressing Asian American disparities in utilization of clinical preventive services: 12-year pooled data from California. *Health Equity*. 2020;4(1):292–303. doi: 10.1089/heq.2020.0008
9. Mok JW, Oh YH, Magge D, Padmanabhan S. Racial disparities of gastric cancer in the USA: an overview of epidemiology, global screening guidelines, and targeted screening in a heterogeneous population. *Gastric Cancer*. 2024;27(3):426–38. doi: 10.1007/s10120-024-01475-9
10. Shah S, Canakis A, Peek RM, Jr., Saumoy M. Endoscopy for gastric cancer screening is cost-effective for Asian Americans in the United States. *Clin Gastroenterol Hepatol*. 2020;18(3):3026–39.
11. Lee RJ, Madan RA, Kim J, Posadas EM, Yu EY. Disparities in cancer care and the Asian American population. *Oncologist*. 2021;26(6):453–60. doi: 10.1002/onco.13748
12. Zack T, Lehman E, Suzgun M, et al. Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study. *Lancet Digit Health*. 2024;6(1):e12–22. doi: 10.1016/S2589-7500(23)00225-X
13. Eckhardt CM, Madjarova SJ, Williams RJ, et al. Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surg Sports Traumatol Arthrosc*. 2023;31(2):376–81. doi: 10.1007/s00167-022-07233-7
14. Zhu S, Wang W, Liu Y. Quite good, but not enough: nationality bias in large language models – a case study of ChatGPT. arXiv. 2024. doi: 10.48550/arXiv.2405.06996
15. United States Census Bureau. *Asian American, Native Hawaiian and Pacific Islander Heritage Month: May 2023*. Census.gov.; 2023. Available from: https://www.census.gov/newsroom/facts-for-features/2023/asian-american-pacific-islander.html#:~:text=The%20estimated%20number%20of%20people,and%20Japanese%20(1.6%20million) [Accessed October 1, 2024].
16. Lee S, Martinez G, Ma GX, et al. Barriers to health care access in 13 Asian American communities. *Am J Health Behav*. 2010;34(1):21–30. doi: 10.5993/ajhb.34.1.3
17. Goodman RS, Patrinely JR, Stone CA, Jr, et al. Accuracy and reliability of chatbot responses to physician questions. *JAMA Netw Open*. 2023;6(10):e2336483. doi: 10.1001/jamanetworkopen.2023.36483
18. Li Y, Li Z, Zhang K, Dan R, Jiang S, Zhang Y. ChatDoctor: a medical chat model fine-tuned on a large language model meta-AI (LLaMA) using medical domain knowledge. *Cureus*. 2023;15(6):e40895. doi: 10.7759/cureus.40895
19. Breaking the token limit: how to work with large amounts of text in chatgpt. *Medium*. February 12, 2023. Available from: https://marco-gonzalez.medium.com/breaking-the-token-limit-how-to-work-with-large-amounts-of-text-in-chatgpt-da18c798d882 [Accessed October 1, 2024].
20. Sai M. Open-source large language models and their maximum input token capacities: a comprehensive analysis. *Medium*. March 8, 2025. Available from: https://medium.com/@saimoguloju2/open-source-large-language-models-and-their-maximum-input-token-capacities-a-comprehensive-7d663347ec61 [Accessed October 1, 2024].
21. Kowsher M, Panditi R, Prottasha NJ, Bhat P, Bairagi AK, Arefin MS. Token trails: navigating contextual depths in conversational AI with ChatLLM. In: Rapp A, Di Caro L, Meziane F, Sugumaran V, eds. *Natural Language Processing and Information Systems*. Lecture Notes in Computer Science. Cham: Springer; 2024:56–67. doi: 10.1007/978-3-031-70242-6_6
22. Davis J, Van Bulck L, Durieux BN, Lindvall C. The temperature feature of CHATGPT: modifying creativity for clinical research. *JMIR Hum Factors*. 2024;11:e53559. doi: 10.2196/53559
23. Yan C. Understanding temperature setting in generative AI models. *Medium*. June 16, 2024. Available from: https://chrisyandata.medium.com/understanding-temperature-setting-in-generative-ai-models-be65489b82fd [Accessed October 1, 2024].
24. Gupta A, Waldron A. Sharing Google's med-palm 2 medical large language model, or LLM. *Google Cloud Blog*. April 13, 2023. Available from: https://cloud.google.com/blog/topics/healthcare-life-sciences/sharing-google-med-palm-2-medical-large-language-model [Accessed October 1, 2024].

Appendix

Supplemental Table 1. ChatGPT–4o Responses in English

Model	Language	Query	Response
ChatGPT-4o	**English**	**What is Gastric Cancer?**	Gastric cancer, also known as stomach cancer, is a type of cancer that starts in the stomach. It can develop in any part of the stomach and may spread to other parts of the body, including the esophagus, liver, lungs, and lymph nodes. Here are some key points about gastric cancer: ### Types of Gastric Cancer 1. **Adenocarcinoma**: This is the most common type, starting in the glandular cells of the stomach lining. 2. **Lymphoma**: Cancer that starts in the immune system tissues within the stomach. 3. **Gastrointestinal Stromal Tumor (GIST)**: A rare type that starts in special cells in the stomach wall. 4. **Carcinoid Tumor**: Cancer that starts in hormone-producing cells of the stomach. ### Risk Factors - **Helicobacter pylori infection**: A common bacterial infection that can cause stomach ulcers and increase cancer risk. - **Diet**: High intake of smoked, salted, or pickled foods and low intake of fruits and vegetables. - **Smoking**: Increases the risk of stomach cancer. - **Genetics**: Family history of gastric cancer or certain inherited genetic conditions. - **Other factors**: Age (more common in older adults), chronic gastritis, pernicious anemia, and previous stomach surgery. ### Symptoms - Early-stage gastric cancer often causes no symptoms or only vague symptoms. As the disease progresses, symptoms may include: - Indigestion or stomach discomfort - Bloating after meals - Nausea and vomiting - Loss of appetite - Unintended weight loss - Stomach pain - Difficulty swallowing (if the tumor is near the esophagus) - Blood in stool or vomit ### Diagnosis Diagnosis typically involves a combination of: - **Endoscopy**: A procedure where a flexible tube with a camera is inserted through the mouth to view the stomach lining. - **Biopsy**: Taking a tissue sample during endoscopy to check for cancer cells. - **Imaging tests**: Such as CT scans, PET scans, and X-rays to determine the extent of the cancer. ### Treatment Treatment depends on the stage and location of the cancer and the patient's overall health: - **Surgery**: To remove part or all of the stomach (gastrectomy) and possibly nearby lymph nodes. - **Chemotherapy**: To kill cancer cells or shrink tumors before surgery. - **Radiation therapy**: To destroy cancer cells, often used in combination with chemotherapy. - **Targeted therapy**: Drugs that target specific aspects of cancer cells, such as HER2 protein. - **Immunotherapy**: Uses the body's immune system to fight cancer. ### Prognosis The outlook for gastric cancer depends on various factors, including the stage at diagnosis and the patient's overall health. Early detection and treatment improve the chances of a better outcome. If you have more specific concerns or symptoms, consulting a healthcare professional for a thorough evaluation and appropriate testing is essential.