communications medicine

A Nature Portfolio journal

Article

https://doi.org/10.1038/s43856-025-00802-0

Large language model comparisons between English and Chinese query performance for cardiovascular prevention

Hongwei Ji[1,2,3,17], Xiaofei Wang[4,17], Ching-Hui Sia[5,6], Jonathan Yap[7,8], Soo Teik Lim[7], Andie Hartanto Djohan[5,6], Yaowei Chang[9], Ning Zhang[2], Mengqi Guo[2], Fuhai Li[2], Zhi Wei Lim[10,11], Ya Xing Wang[1,3], Bin Sheng[12], Tien Yin Wong[1,3,13], Susan Cheng[14,18], Khung Keong Yeo[7,18] & Yih-Chung Tham[11,13,15,16,18]

Abstract

Background Large language model (LLM) offer promise in addressing layperson queries related to cardiovascular disease (CVD) prevention. However, the accuracy and consistency of information provided by current general LLMs remain unclear.

Methods We evaluated capabilities of BARD (Google's bidirectional language model for semantic understanding), ChatGPT-3.5, ChatGPT-4.0 (OpenAI's conversational models for generating human-like text) and ERNIE (Baidu's knowledge-enhanced language model for context understanding) in addressing CVD prevention queries in English and Chinese. 75 CVD prevention questions were posed to each LLM. The primary outcome was the accuracy of responses (rated as appropriate, borderline, inappropriate).

Results For English prompts, the chatbots' appropriate ratings are as follows: BARD at 88.0%, ChatGPT-3.5 at 92.0%, and ChatGPT-4.0 at 97.3%. All models demonstrate temporal improvement in initially suboptimal responses, with BARD and ChatGPT-3.5 each improving by 67% (6/9 and 4/6), and ChatGPT-4.0 achieving a 100% (2/2) improvement rate. Both BARD and ChatGPT-4.0 outperform ChatGPT-3.5 in recognizing the correctness of their responses. For Chinese prompts, the "appropriate" ratings are: ERNIE at 84.0%, ChatGPT-3.5 at 88.0%, and ChatGPT-4.0 at 85.3%. However, ERNIE outperform ChatGPT-3.5 and ChatGPT-4.0 in temporal improvement and self-awareness of correctness.

Conclusions For CVD prevention queries in English, ChatGPT-4.0 outperforms other LLMs in generating appropriate responses, temporal improvement, and self-awareness. The LLMs' performance drops slightly for Chinese queries, reflecting potential language bias in these LLMs. Given growing availability and accessibility of LLM chatbots, regular and rigorous evaluations are essential to thoroughly assess the quality and limitations of the medical information they provide across widely spoken languages.

Plain Language Summary

Recently there has been an increase in the use of large language model (LLM) chatbots by patients seeking medical information. However, the accuracy of information provided by LLMs across different languages remain unclear. This study aimed to evaluate the performance of popular LLM chatbots, such as BARD, ChatGPT-3.5, ChatGPT-4.0, and ERNIE, in answering cardiovascular disease prevention questions in both English and Chinese. We tested these models with 75 questions each, focusing on the accuracy of their responses and their ability to improve over time. The results showed that ChatGPT-4 provided the most accurate answers in English and demonstrated the best improvement over time. In Chinese, ERNIE performed better in improving its responses over time. This research highlights the need for ongoing evaluations to ensure the spread of reliable health information by LLMs across diverse languages.

Large language models (LLMs) consist of a neural network with typically billions of parameters, trained on large quantities of available text including those relevant to health care[1–4]. With the rapid development of LLMs, along with mass scale of data available for training, it is now possible for LLM to provide relatively appropriate answers and responding in a human-like manner when prompted with health-related queries[5–7]. LLM Chatbots such as BARD (Google's bidirectional language model for semantic understanding) (https://bard.google.com/), ERNIE (Baidu's knowledge-enhanced

A full list of affiliations appears at the end of the paper. e-mail: thamyc@nus.edu.sg

language model for context understanding) (https://wenxin.baidu.com/ernie3), and ChatGPT (OpenAI's conversational language models for generating human-like text) (https://chat.openai.com/chat) are now readily accessible by public users[8]. In a time where the internet is becoming a go-to source for healthcare information[9], these widely available and human-like LLM chatbots are set to serve as a resource for a broad range of individuals and communities. However, it remains uncertain how consistently these chatbots provide accurate and evidence-based information.

Cardiovascular disease (CVD) prevention is a topic with extensive evidence-based information. Given the expansive burden of CVD globally[10], LLM chatbots could assist in reducing associated inequities through broader access to high-quality health information on CVD prevention[11,12]. Early exploratory analyses showed that ChatGPT-3.5 is able to address CVD prevention queries with an appropriateness level of up to 88%[6,8]. Beyond ChatGPT-3.5, the relative performances of other common LLM Chatbots in this context have yet been evaluated[13–19].

As LLMs undergo refinement through user interactions and feedback, updated data, and underlying algorithm updates, they could improve over time. Additionally, LLM has the ability to mitigate hallucinations and false claims through self-checking[20]. Therefore, in addition to assessing how different LLM chatbots compare to each other in terms of appropriately answer queries, it is also important to gauge their potential ability to identify and rectify initial inappropriate responses through self-checking and to evaluate the temporal improvement of these models. Overall, understanding the extent to which some LLM chatbots are able to deliver reliable medical information while minimizing the spread of medical misinformation requires robust evaluation.

In this study, we conduct a rigorous evaluation of widely accessible LLM chatbots, assessing their accuracy, temporal improvement, and self-checking capabilities in responding to CVD prevention queries in two predominant languages - English and Chinese. The results indicate that ChatGPT-4 delivers the most accurate answers in English and exhibits the greatest improvement over time, while ERNIE shows better performance in enhancing its responses in Chinese.

Method

Large language models and Chatbots

To ensure a comprehensive evaluation that includes common and popular LLM-chatbots, we included four LLM Chatbots in our study: (1) ChatGPT-3.5 by OpenAI; (2) ChatGPT-4.0 by OpenAI; (3) BARD by Google; and (4) ERNIE by Baidu. The evaluation of English prompts, involved ChatGPT 3.5, ChatGPT 4, and BARD; for Chinese prompts, the evaluation involved ChatGPT 3.5, ChatGPT 4, and ERNIE (Supplementary Table S1). We used the models with their default configurations and temperature settings (temperature of 0.7 for ChatGPT-3.5 and -4.0; 0.5 for BARD; and 0.8 for ERNIE).We did not make any adjustments to these parameters during our analysis. This study does not include human or animal subjects, and so the need for ethical review was waived.

Question generation and chatbot response evaluation

The American College of Cardiology and American Heart Association provide guidelines and recommendations for CVD preventionss[6,21,22]. These guidelines encompass information on risk factors, diagnostic tests, and treatment options, as well as patient education and self-management strategies. Drawing from key topics within these guidelines, we involved 2 experienced attending-level cardiologists (YWC, HWJ) to generate questions related to CVD prevention, framing them similarly to how patients would inquire with physicians to ensure relevance and comprehensibility from a patient's perspective (Supplementary Table S2). This patient-centered and guideline-based approach yielded a final set of 300 questions covering domains such as biomarker, medication information, dyslipidemia, hypertension, diet counseling, diabetes mellitus and/or chronic kidney disease, secondary prevention, prevention strategy, inflammation, exercise counseling, obesity, and tobacco treatment. We then translated these questions into Chinese, ensuring the appropriate use of conventional units (e.g., mg/dL) and international units (e.g., mmol/L) in the English and Chinese versions respectively

Figure 1 depicts the comprehensive study design employed in our research. To enhance the reliability and minimize bias, a random sampling approach was implemented. Our target sample size was determined to be 225 (75 per chatbot), aiming for a 90% power to detect an effect size of 5% within a general linear model framework. Consequently, we randomly selected 75 questions from the original pool of 300 questions (Supplementary Table S3). Between the dates of 24th April and 9th May, separate query sessions were conducted for ChatGPT-3.5 (both English and Chinese versions), ChatGPT-4.0 (both English and Chinese versions), BARD (English only), and ERNIE (Chinese only). Each chatbot was utilized to respond to 75 prompts, with each prompt being posed once on the interface during the respective query session (Supplementary Fig. S1). This process yielded a total of 75 responses per chatbot. Hence, the final sample consisted of 75:75:75 responses generated by the three LLM-Chatbots, for both the English and Chinese sections of the study. The responses were independently evaluated by two panels of cardiologists from Singapore and China, each panel comprising cardiologists with a minimum of five years of practice in cardiology. One panel assessed the English responses, and the other evaluated the Chinese responses, ensuring proficiency in the respective languages.

Blinding and randomly ordered assessment

To ensure the graders were unable to distinguish the origin of the response among different LLM Chatbots, we manually concealed any chatbot-specific features. These features included phrases such as "I am not a doctor" by GPT-4, which could indicate the use of a specific model. (Supplementary Table S4). As shown in Fig. 1, the evaluation was conducted in a blinded and randomly ordered manner. Specifically, the responses from three chatbots were randomly shuffled within the question set. The responses from three chatbots were randomly assigned to 3 rounds, in a 1:1:1 ratio, for blinded assessment by three cardiologists, with a 48-hour wash-out interval in between rounds so as to mitigate recency bias.

Accuracy evaluation

The primary outcome in this study was the performance in responding to primary CVD prevention questions. Specifically, we used a two-step approach to evaluate the responses. In the first step, the panel of three cardiologists (CHS, JY, AHD for English; ZN, GMQ, LFH for Chinese), reviewed all LLM Chatbot generated responses and graded the responses as: "appropriate", "borderline" or "inappropriate", in relation to expert consensus and guidelines[21–23]. Specifically, the response was graded as appropriate when there were no inaccuracies; borderline when there were potential factual inaccuracies but still unlikely to mislead the average patient or cause harm; and inappropriate when the response consisted of unacceptable inaccuracies that would likely mislead the average patient and cause harm. In the second step, we utilized a majority consensus approach, wherein the final rating for each chatbot response was based on the most common rating graded amongst the three graders. In scenarios where majority consensus could not be achieved among the three graders (i.e., each grader provided a different rating), we additionally sought adjudication from a senior cardiologist (STL) to finalize the rating.

Evaluations of temporal improvement and self-checking capabilities

The secondary outcomes involved the evaluation of temporal improvement and self-checking capabilities of the LLM Chatbots. Therefore, we targeted prompts that initially yielded suboptimal responses to create a controlled environment conducive to evaluating the self-awareness capabilities of the LLMs in identifying and rectifying errors, which also helps establish a baseline of common issues and errors. Specifically, to evaluate temporal improvements, we implemented an experimental process that allowed responses initially deemed as "borderline" or "inappropriate" to be given a subsequent opportunity for refinement (2 months after original baseline

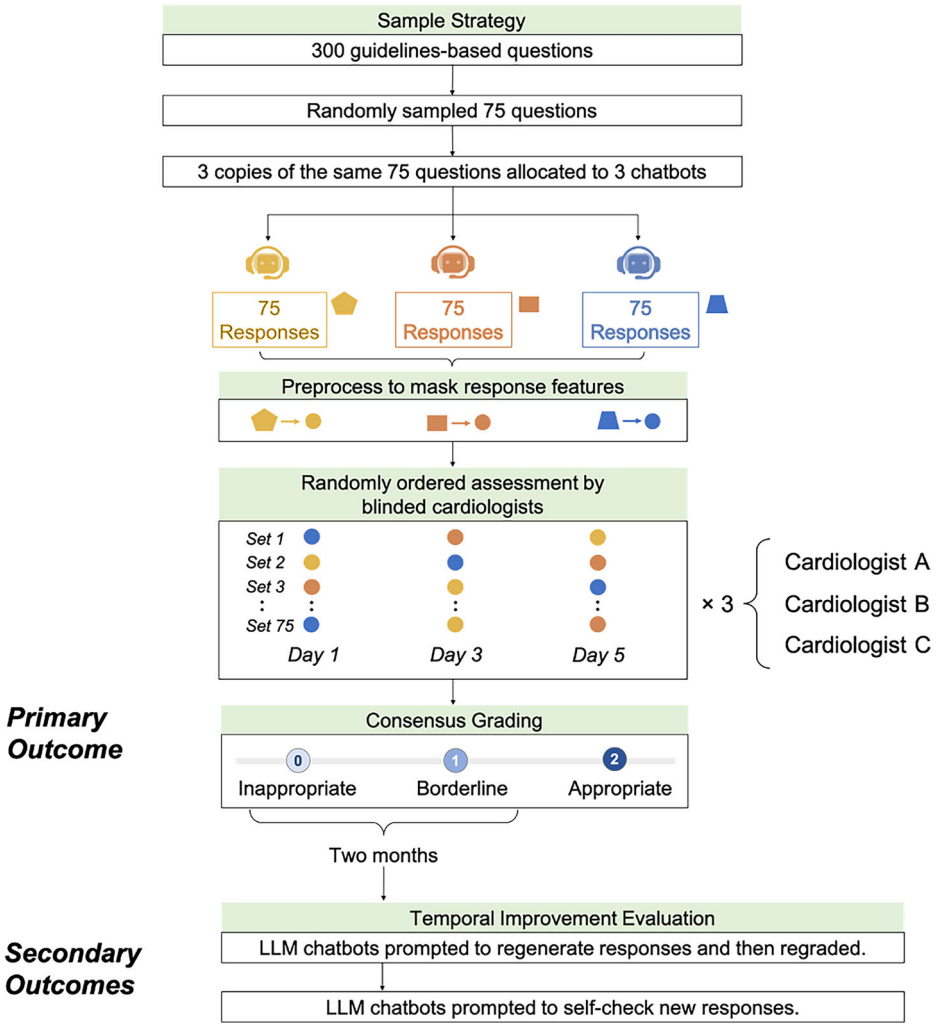

Fig. 1 | Study design. Sample strategy, preprocess, randomly-ordered assessment by blinded cardiologist. 75 questions were randomly selected from the original pool of 300 questions. Each chatbot was utilized to respond to 75 prompts, with each prompt being posed once on the interface during the respective query session. The evaluation was conducted in a blinded and randomly ordered manner. Specifically, the responses from three chatbots were randomly shuffled within the question set. The responses from three chatbots were randomly assigned to 3 rounds, in a 1:1:1 ratio, for blinded assessment by three cardiologists, with a 48-hour wash-out interval in between rounds so as to mitigate recency bias.

response). These regenerated responses then underwent another round of expert evaluation. To evaluate the LLMs self-checking capabilities, we further prompted the chatbots with "please check if the above answer is correct" to discern whether the chatbots could self-check for correctness or further improve the quality of their original incorrect response, when prompted.

Statistics and reproducibility

Data were presented as mean (SD), and n (%) as appropriate. Kruskal-Wallis Rank Sum Test and Mann–Whitney U test (for pairwise) were used to compare the differences in sum score of responses across the three LLM-Chatbots. To compare the proportions of 'appropriate,' 'borderline,' and 'inappropriate' ratings across the LLM Chatbots, a two-tailed Pearson's $\chi 2$ test was conducted. All analyses were performed using R version 4.2.1 (R Foundation for Statistical Computing). A two-sided P value of <0.05 was considered statistically significant.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Results

Responses from LLM-Chatbots

Table 1 displays the length of the LLM-Chatbots' responses to the 75 selected CVD prevention queries. The mean (standard deviation, [SD]) of the word count of English response was 209.08 (70.82) for Google Bard, 165.01 (55.60) for ChatGPT-3.5, and 213.28 (83.67) for ChatGPT-4.0. The mean (SD) word count of Chinese response was 299.68 (119.10) for Baidu ERNIE, 320.44 (100.54) for ChatGPT-3.5, and 405.73 (134.86) for ChatGPT-4.0.

Performance of LLM-Chatbots

Table 2 displays the performance of different LLM-Chatbots. LLM-chatbot performed generally better with English prompts than with Chinese prompts (Pearson's chi-squared test, $P = 0.022$). Specifically, for English prompts, BARD (sum score of 5.40), ChatGPT-3.5 (score of 5.45), and ChatGPT-4.0 (score 5.65) demonstrated similar sum score (P values ranging from $P = 0.057$ to $P = 0.74$). When comparing proportions of 'appropriate' rating, ChatGPT-4.0 had 97.3% of 'appropriate' rating, compared to 92% in ChatGPT-3.5 (Pearson's chi-squared test, $P = 0.24$) and 88% in Google Bard (Pearson's chi-squared test, $P = 0.021$). For Chinese prompts, ChatGPT3.5 had a higher sum score (5.25), followed by ChatGPT-4.0 (5.07), and Ernie (4.99). However, the differences were not statistically significant (P-values ranging from $P = 0.34$ to $P = 0.78$) Similarly, ChatGPT-3.5 had higher proportion of 'appropriate rating' for Chinese prompts (88.0%), compared to ChatGPT-4.0 (85.3%) and ERNIE (84.0%), but the differences were not statistically significant (P values ranging from $P = 0.71$ to $P = 0.96$).

Performance across CVD prevention domains

Figure 2 illustrated the "appropriate" ratings across different CVD prevention domains. Remarkably, ChatGPT-4.0 consistently performed well in most domains, with a 100% "appropriate" rating in "dyslipidemia", "lifestyle", "biomarker and inflammation", and "DM and CKD" domains

Table 1 | Overview of response length from LLM-Chatbots to cardiovascular disease prevention queries

English Response Length	BARD	ChatGPT 3.5	ChatGPT 4	$P_{BARD\ vs.\ 3.5}$	$P_{BARD\ vs.\ 4}$	$P_{3.5\ vs.\ 4}$	P_{ANOVA}
Words, mean (SD)	209.08 (70.82)	165.01 (55.60)	213.28 (83.67)	<0.001	0.74	<0.001	<0.001
Chinese Response Length	**ERNIE**	**ChatGPT 3.5**	**ChatGPT 4**	$P_{ERNIE\ vs.\ 3.5}$	$P_{ERNIE\ vs.\ 4}$	$P_{3.5\ vs.\ 4}$	P_{ANOVA}
Words, mean (SD)	299.68 (119.10)	320.44 (100.54)	405.73 (134.86)	0.25	<0.001	<0.001	<0.001

The p-values in the table represent the following comparisons: $P_{BARD\ vs.\ 3.5}$: Comparison between Google Bard and ChatGPT-3.5. $P_{BARD\ vs.\ 4}$: Comparison between Google Bard and ChatGPT-4. $P_{3.5\ vs.\ 4}$: Comparison between ChatGPT-3.5 and ChatGPT-4. P_{ANOVA}: P-value from the ANOVA test comparing all three models (Google Bard, ChatGPT-3.5, and ChatGPT-4).
SD standard deviation.

Table 2 | Performance of LLM-Chatbots in addressing questions with english prompts

English Prompts	BARD	ChatGPT 3.5	ChatGPT 4	$P_{BARD\ vs.\ 3.5}$	$P_{BARD\ vs.\ 4}$	$P_{3.5\ vs.\ 4}$
Sum Score, mean (SD)[a]	5.40 (0.93)	5.45 (1.06)	5.65 (0.67)	0.74	0.057	0.16
Appropriate, *n* %	66 (88.0)	69 (92.0)	73 (97.3)	0.33	0.021	0.24
Borderline, *n* %	9 (12.0)	5 (6.7)	1 (1.3)			
Inappropriate, *n* %	0 (0.0)	1 (1.3)	1 (1.3)			
Chinese Prompts	**ERNIE**	**ChatGPT 3.5**	**ChatGPT 4**	$P_{ERNIE\ vs.\ 3.5}$	$P_{ERNIE\ vs.\ 4}$	$P_{3.5\ vs.\ 4}$
Sum Score, mean (SD)[a]	4.99 (1.85)	5.25 (1.62)	5.07 (1.74)	0.34	0.78	0.49
Appropriate, *n* %	63 (84.0)	66 (88.0)	64 (85.3)	0.71	0.96	0.85
Borderline, *n* %	3 (4.0)	3 (4.0)	3 (4.0)			
Inappropriate, *n* %	9 (12.0)	6 (8.0)	8 (10.7)			

[a]For gradings from three cardiologists, "Appropriate" was assigned with score 2, "Borderline" was assigned with score 1, and "Inappropriate" was assigned with score 0. The p-values in the table represent the following comparisons: $P_{BARD\ vs.\ 3.5}$: Comparison between Google Bard and ChatGPT-3.5. $P_{BARD\ vs.\ 4}$: Comparison between Google Bard and ChatGPT-4. $P_{3.5\ vs.\ 4}$: Comparison between ChatGPT-3.5 and ChatGPT-4.
SD standard deviation.

(Supplementary Table S5). However, BARD showed suboptimal performance compared to ChatGPT4.0 and ChatGPT-3.5, particularly in the "lifestyle" domain, where ChatGPT-4.0 and ChatGPT-3.5 achieved 100% "appropriate" ratings compared to 73.3% for BARD ($P = 0.012$).

Figure 3 showed results for Chinese prompts. All three LLM-Chatbots performed well in the "lifestyle" domain, with 100% "appropriate" ratings (Supplementary Table S6). However, variations in performance were observed across other domains. ChatGP-4.0 and ChatGPT-3.5 performed better in "biomarker and inflammation" and "prevention strategy" domains, whereas ERNIE performed better in "DM and CKD" and "prevention strategy" domains.

Progressive improvement and self-checking ability

Table 3 highlights the progressive improvement of LLM-Chatbots over 2 months and their ability to self-check when prompted. Overall, all LLM-Chatbots exhibited substantial improvements in rectifying initial suboptimal responses with their updated iterations. ChatGPT-3.5 improved 66.7% of its suboptimal responses (4 out of 6), ChatGPT-4.0 improved 100% (2 out of 2), and Google Bard improved 66.7% (6 out of 9). When prompted with "please check if the above answer is correct", ChatGPT-4.0 identified the correctness of 100% of its answers (2 out of 2), BARD identified 77.8% (7 out of 9), and ChatGPT-3.5 identified 16.7% (1 out of 6). For Chinese prompts, ERNIE improved 91.6% (11 out of 12), ChatGPT-4.0 improved 54.5% (6 out of 11), and ChatGPT-3.5 improved 22.2% (2 out of 9) of their initial suboptimal responses. When prompted to check their own answer, ERNIE identified the correctness of 91.6% of its answers (11 out of 12), ChatGPT-4.0 identified 45.4% (5 out of 6), and ChatGPT-3.5 identified 11.1%(1 out of 9).

Discussion

Our study showcases a head-to-head comparison of LLM-chatbot performance in addressing CVD prevention questions using 2 predominant languages – English and Chinese. Notably, LLM-chatbots exhibited significant disparities in performance across languages, performing generally better with English prompts than with Chinese prompts. ChatGPT-4.0 outperformed ChatGPT-3.5 and Bard for English prompts while ChatGPT-3.5 outperformed ChatGPT-4.0 and ERNIE for Chinese prompts. When evaluating for temporal improvement and self-checking capabilities, ChatGPT-4.0 and ERNIE exhibited substantial improvements in rectifying initial suboptimal responses with their updated iterations for English and Chinese prompts, respectively. Our study findings highlight the promising capabilities of LLM-Chatbots in addressing inquiries related to CVD prevention and its potential for future advancements in this field.

This study has important implications for CVD prevention. Individuals with health concerns have become increasingly engaged consumers of publicly available health information[24] – beginning with the advent of the digital age and, then, even more so during the recent growth in telehealth care programs and augmentation of internet search functions[25]. The traditional model of patients seeking and gaining information from their primary care providers has been shown historically to enhance knowledge and understanding of cardiovascular risk factors, healthy behaviors, and preventive measures related to cardiovascular health[26]. However, the sole reliance on primary care practitioners to improve population's cardiovascular literacy poses inherent limitations such as geographical and resource disparities and time constraints[27–29], especially when considering the poor access to health care in underserved population[30]. LLM-Chatbots offer promising potential in delivering accurate knowledge and information to bridge these gaps. In this regard, our study provides valuable evidence regarding the utility of appropriate LLM-Chatbots in promoting health literacy in terms of CVD prevention.

Moreover, in the context of responding to CVD prevention queries, this study represents the first investigation comparing the performance of chatbots when prompted with Chinese[17–19], a language used by ~20% of the global population. Baidu's ERNIE, tailored for Chinese linguistic nuances[31], displayed inherent strengthswhen compared to ChatGPT 3.5 and ChatGPT 4 in distinct CVD prevention areas. Notably, ChatGPT often misidentified specific drug brand names in Chinese, mistakenly linking '诺欣妥' (nuoxintuo, the Chinese moniker for sacubitril-valsartan) to unrelated

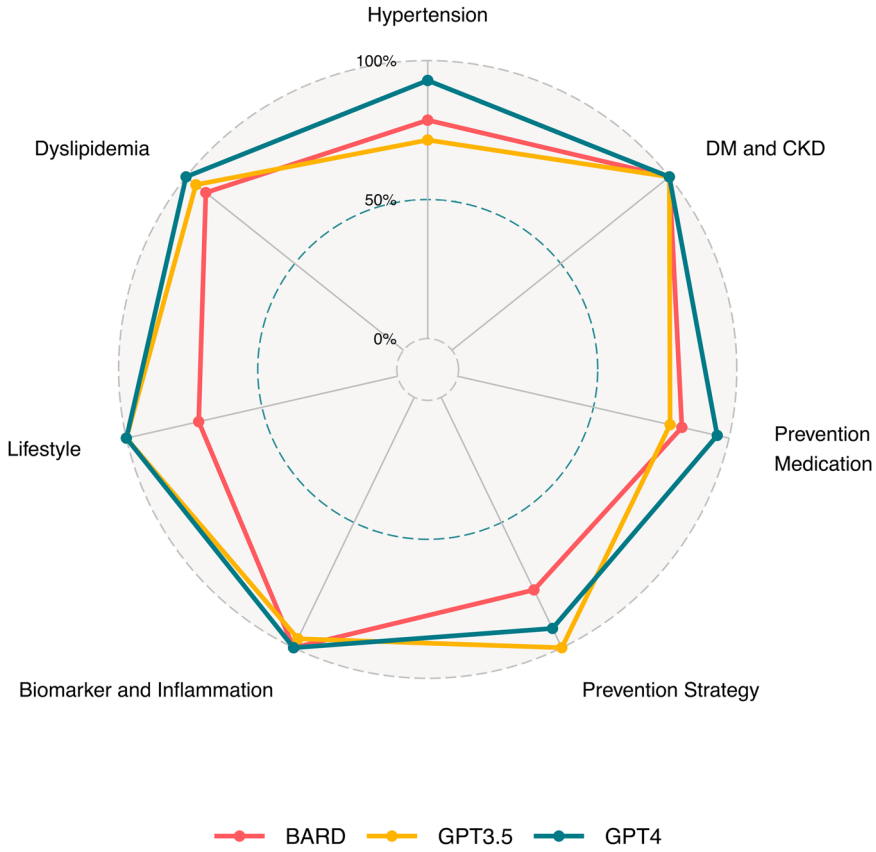

Fig. 2 | Responses graded "Appropriate" across subject matter domains of cardiovascular prevention using English prompts. This spider plot showed large language model (LLM) with English prompts. Points indicate the percentage of responses from the respective LLM that were graded as appropriate. Lines indicate the performance of different LLMs, with each color representing a different model. Dotted lines indicate reference points at 0%, 50%, and 100%. Gray lines to the center indicate that points on these lines are for the same LLM. DM diabetes mellitus, CKD chronic kidney disease.

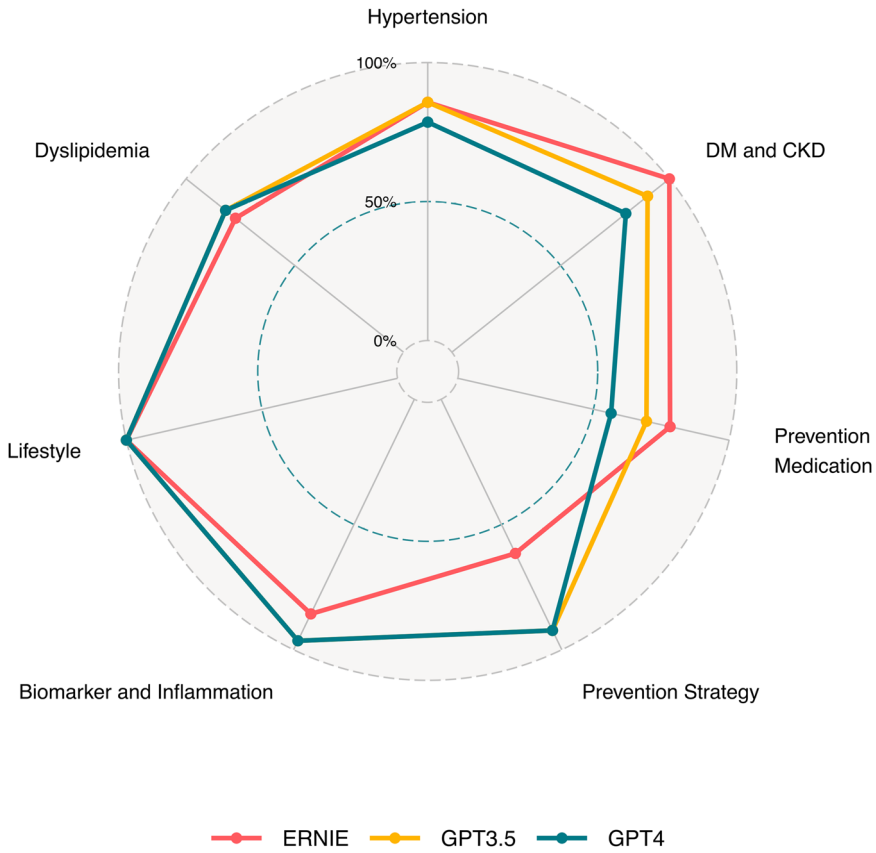

Fig. 3 | Responses graded "Appropriate" across subject matter domains of cardiovascular prevention using Chinese prompts. This spider plot showed large language model (LLM) with Chinese prompts. Points indicate the percentage of responses from the respective LLM that were graded as appropriate. Lines indicate the performance of different LLMs, with each color representing a different model. Dotted lines indicate reference points at 0%, 50%, and 100%. Gray lines to the center indicate that points on these lines are for the same LLM. DM diabetes mellitus, CKD chronic kidney disease.

Table 3 | Performance of LLM-Chatbots in refining suboptimal responses with updated model iterations (English)

English Prompts	BARD	ChatGPT 3.5	ChatGPT 4
Number of Suboptimal Responses, *n*	9	6	2
Temporal Improvement, *n* (%)[a]	6 (66.7)	4 (66.7)	2 (100)
Self-check, *n* (%)[b]	7 (77.8)	1 (16.7)	2 (100)
Chinese Prompts	**ERNIE**	**ChatGPT 3.5**	**ChatGPT 4**
Number of Suboptimal Responses, *n*	12	9	11
Temporal Improvement, *n* (%)[a]	11 (91.6)	2 (22.2)	6 (54.5)
Self-check, *n* (%)[b]	11 (91.6)	1 (11.1)	5 (45.4)

[a]To evaluate the models' temporal improvement over the study period, suboptimal responses which included "borderline" and "inappropriate" responses were updated using the latest iteration of LLM (26th June and 10th July), with new responses assessed by the cardiologist graders.
[b]To test the performance of self-check, "please check if above answer is correct" was entered as a follow-up prompt. Successful self-check was defined as either recognizing whether the response is correct or additionally providing an appropriate response.
LLM large language model, *SD* standard deviation

drugs like Norspan and Norinyl. This suggests a possible over-reliance on transliteration techniques for Chinese drug names. Although ERNIE excelled at drug name recognition, its overall competency across domains still fell short in comparison to ChatGPTs for Chinese queries. While ERNIE was developed with the goal of improving access to health information[32], especially in Chinese speaking regions, our findings however indicate that it did not surpass ChatGPT in performance. These findings may suggest that current language specific LLM may not be as well and broadly trained as generic LLM such as ChatGPT. The performance disparities observed likely stem from the quality and availability of training datasets. This distinction is particularly evident when juxtaposing English and Chinese LLM capabilities, given the varying quality of guideline-based CVD prevention resources across the languages. Despite the initial postulation, our findings are noteworthy in revealing that Chinese-specific LLM performed inferiorly compared to generic LLMs like ChatGPT-4.0. This disparity may suggest that despite being tailored to the Chinese language, the current Chinese-specific LLMs may not have been trained as broadly as the generic, English dominant LLMs[33]. In addition, our findings demonstrated variability in response lengths across different LLMs. While longer responses may suggest a more comprehensive understanding of the query topic, this increased verbosity did not consistently lead to higher accuracy rates. For instance, among Chinese responses, a mean response length of 299 words was associated with an accuracy of 84%, while a length of 405 words corresponded to a just slightly higher accuracy of 85.3%. Nevertheless, the impact of response length on perceived accuracy warrants further evaluation.

Our assessment covered both the chatbots' initial factual accuracy and their adeptness at refining suboptimal responses over time. Recent updates to ChatGPT 3.5, ChatGPT 4, and BARD have shown marked improvements, transitioning their responses from "inappropriate" or "borderline" to "appropriate" ones. Our findings were consistent with Johnson et. al. paper[34], which reported a significant improvement in accuracy scores over a 2-week period between evaluations. Collectively, these exemplify the rapidly advancing nature of LLMs and its boundless potential moving forward. Additionally, we further examined the chatbots' self-awareness of correctness by instructing them to review their own responses. Interestingly, ChatGPT 3.5, even in its updated form, identified the correctness of only 1 out of 6 of its own responses. This indicates that, even when explicitly prompted, LLM-Chatbots might continue to relay inaccurate information. Moreover, the gaps in ability to improve over time are related not only to availability and quality of training data but in the availability and quality of the continued interaction and feedback data. Thus, it is likely that the LLM chatbots that will demonstrate substantial improvements in performance over time are those that garner the most attention to ongoing technical improvement but also the most attention in terms of user feedback. Such that we will likely see not only improvement in LLM chatbot performance over time but also increasing gap between high performers and low performers. This will obligate ongoing continuation of comparison studies of this type to understand the magnitude, nature, and temporal trends in these gaps. Regarding Chinese prompts, ERNIE displayed significant improvements in refining suboptimal responses, effectively addressing 11 out of 12 cases. Furthermore, ERNIE demonstrated a significant capability to self-aware correctness, accurately assessing the accuracy of its responses in 11 out of 12 cases. Considering the notable evolutions of LLMs, it should be noted that ChatGPT has undergone a series of more than ten updates[35], whereas Baidu ERNIE has also undergone substantial and pivotal updates[36]. In light of this, the observed disparities of LLMs' temporal improvements should be plausibly attributed to divergent magnitudes and velocities characterizing the updates received by each model. Capitalizing on the promising ability of chatbots to self-check accuracy may entail user adjustments in interaction patterns or enhancements in the chatbot's built-in algorithm checks[37,38], especially concerning medical queries.

In the process of conducting our study, several noteworthy strengths emerged that we believe contribute significantly to the value and reliability of our findings. First, where many studies focused primarily on ChatGPT3.5, we expanded our scope to include ChatGPT-3.5, ChatGPT-4.0, Google Bard, and Baidu ERNIE (Chinese). This comprehensive approach provides a broader understanding of chatbot capabilities in CVD-related patient interactions. Second, our study involved systematic masking, randomization, and a wash-out period between grading sets. Each assessment was meticulously conducted by three seasoned cardiologists, with a consensus approach guiding the establishment of the ground truth. These measures ensured our study's robustness. Third, with our team's multilingual expertise, we could compare chatbot performance in both English and Chinese, offering a unique angle on AI-driven medical communications across major languages. Additionally, beyond assessing a chatbot's factual accuracy, we scrutinized its response evolution and introduced a procedure to prompt self-assessment, highlighting potential avenues for improving AI responses in medical contexts. There are also limitations that may merit further consideration. First, although we generated the questions with a guideline-based approach, they only represent a small part of questions in terms of CVD prevention. Though we compared the responses of five LLMs under consistent conditions to ensure the impact of stochasticity was consistent across the selected LLMs, the impact of stochastic responses may not be eliminated completely. Thus, the generalizability of our findings to the entire spectrum of CVD prevention questions may be limited. Second, though we tested models' temporal improvements, most responses were generated using chatbots between 24th April and 9th May 2023. As the LLM-Chatbots evolve at a unprecedent speed, more continuous research is needed to accommodate updated LLM iterations and other emerging LLMs such as Meta's LLaMA and Anthropic's Claude. Third, to reduce the bias from language proficiency, English part and Chinese part were assessed by independent panel of cardiologists, leading to varying guideline interpretations in respective regions. For example, Entresto was approved for treatment of hypertension in China and Japan but not in United States and Singapore. Thus, the any direct comparisons between performances of the chatbots in response to English and Chinese prompts should be interpreted with caution. Fourth, our findings indicate comparable performances between the chatbots, suggesting a smaller effect size than anticipated. This implies that our initial effect size estimation during the study design phase, set at 0.05, might have been optimistic, resulting in a potential underestimation of the ideal sample size.

In conclusion, ChatGPT-4.0 excels in responding to English-language queries related to CVD prevention, with a high accuracy rate of 97.3%. In contrast, all LLM Chatbots demonstrated moderate performance for Chinese-language queries, with accuracy rates ranging from 84% to 88%. Considering the increasing accessibility of LLM Chatbots, they offer

promising avenues for enhancing health literacy, particularly among underserved communities. Continuous comparative evaluations assessments are crucial to delve deeper into the quality and limitations of the medical information disseminated by these chatbots across common languages.

Data availability

Source data for Figs. 2 and 3 can be found in Supplementary Table S5 and Supplementary Table S6, respectively. All data are available from the corresponding author (or other sources, as applicable) on reasonable request.

Received: 13 December 2023; Accepted: 11 March 2025;
Published online: 16 May 2025

References

1. Arora, A. & Arora, A. The promise of large language models in health care. *Lancet* **401**, 641 (2023).
2. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
3. Haupt, C. E. & Marks, M. AI-Generated Medical Advice-GPT and Beyond. *JAMA* **329**, 1349–1350 (2023).
4. Lee, P., Bubeck, S. & Petro, J. Benefits, limits, and risks of GPT-4 as an AI Chatbot for medicine. *N. Engl. J. Med.* **388**, 1233–1239 (2023).
5. Sng, G. G. R., Tung, J. Y. M., Lim, D. Y. Z. & Bee, Y. M. Potential and Pitfalls of ChatGPT and natural-language artificial intelligence models for diabetes education. *Diabetes Care* **46**, e103–e105 (2023).
6. Sarraju, A. et al. Appropriateness of cardiovascular disease prevention recommendations obtained from a popular online chat-based artificial intelligence model. *JAMA* **329**, 842–844 (2023).
7. Kassab, J. et al. Assessing the Accuracy of an Online Chat-Based Artificial Intelligence Model in Providing Recommendations on Hypertension Management in Accordance With the 2017 American College of Cardiology/American Heart Association and 2018 European Society of Cardiology/European Society of Hypertension Guidelines. *Hypertension* **80**, e125–e127 (2023).
8. Ayers, J. W. et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
9. Kuehn, B. M. More than one-third of US individuals use the Internet to self-diagnose. *JAMA* **309**, 756–757 (2013).
10. Roth, G. A. et al. Group G-N-JGBoCDW. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. *J. Am. Coll. Cardiol.* **76**, 2982–3021 (2020).
11. LaVeist, T. A. et al. The economic burden of racial, ethnic, and educational health inequities in the US. *JAMA* **329**, 1682–1692 (2023).
12. Coleman, C., Birk, S. & DeVoe, J. Health literacy and systemic racism-using clear communication to reduce health care inequities. *JAMA Intern. Med.* **183**, 753–754 (2023).
13. Azamfirei, R., Kudchadkar, S. R. & Fackler, J. Large language models and the perils of their hallucinations. *Crit. Care* **27**, 120 (2023).
14. Stokel-Walker, C. & Van Noorden, R. What ChatGPT and generative AI mean for science. *Nature* **614**, 214–216 (2023).
15. van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. L. ChatGPT: five priorities for research. *Nature* **614**, 224–226 (2023).
16. Koh, S. J. Q., Yeo, K. K. & Yap, J. J. L. Leveraging ChatGPT to aid patient education on coronary angiogram. *Ann. Acad. Med. Singap.* **52**, 374–377 (2023).
17. Fang, C. et al. How does ChatGPT-4 preform on non-English national medical licensing examination? An evaluation in Chinese language. *PLOS Digit Health* **2**, e0000397 (2023).
18. Cai, Y. et al. MedBench: a large-scale Chinese benchmark for evaluating medical large language models. In *Proc. Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, Vol. 38, 1975 (AAAI, 2024).
19. Junling Liu, P. Z. et al. Benchmarking large language models on CMExam-A Comprehensive Chinese Medical Exam Dataset. *Neural Inform. Process. Syst*. **36**, 52430–52452 (2023).
20. Weng, Y. et al. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023* 2550–2575 (Association for Computational Linguistics, 2023).
21. Arnett, D. K. et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **74**, 1376–1414 (2019).
22. Arnett, D. K. et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **140**, e596–e646 (2019).
23. Grundy, S. M. et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: executive summary: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *J. Am. Coll. Cardiol.* **73**, 3168–3209 (2019).
24. Eysenbach, G. & Diepgen, T. L. Patients looking for information on the Internet and seeking teleadvice: motivation, expectations, and misconceptions as expressed in e-mails sent to physicians. *Arch. Dermatol.* **135**, 151–156 (1999).
25. Tan, S. S. & Goonawardene, N. Internet health information seeking and the patient-physician relationship: a systematic review. *J. Med. Internet Res.* **19**, e9 (2017).
26. Organization WH. Adherence to long-term therapies: evidence for action. *World Health Organization*. https://apps.who.int/iris/handle/10665/42682 (2003).
27. Katzmarzyk, P. T. et al. Weight loss in underserved patients - a cluster-randomized trial. *N. Engl. J. Med.* **383**, 909–918 (2020).
28. Dwyer-Lindgren, L. et al. US County-Level trends in mortality rates for major causes of death, 1980–2014. *JAMA* **316**, 2385–2401 (2016).
29. Lin, C. T. et al. Is patients' perception of time spent with the physician a determinant of ambulatory patient satisfaction? *Arch. Intern. Med.* **161**, 1437–1442 (2001).
30. Munoz, D. et al. Polypill for cardiovascular disease prevention in an underserved population. *N. Engl. J. Med.* **381**, 1114–1123 (2019).
31. Wang, S. et al. ERNIE 3.0 Titan: exploring larger-scale knowledge enhanced pre-training for language understanding and generation. Preprint at https://arxiv.org/abs/2112.12731 (2021).
32. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
33. Tang, T. et al. Not all metrics are guilty: improving NLG evaluation by diversifying references. In *Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, 6596–6610 (Association for Computational Linguistics, 2024).
34. Johnson, D. et al. Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the Chat-GPT model. Preprint at *Res. Sq*. https://doi.org/10.21203/rs.3.rs-2566942/v1 (2023).
35. OpenAI. ChatGPT release notes. OpenAI Help Center. https://help.openai.com/en/articles/6825453-chatgpt-release-notes (2023).
36. Baidu Research. Introducing ERNIE 3.5: Baidu's knowledge-enhanced foundation model takes a giant leap forward. *Baidu Blog*. http://research.baidu.com/Blog/index-view?id=185 (2023).

37. Hu, Z. et al. Unlocking the potential of user feedback: leveraging large language model as user simulators to enhance dialogue system. Preprint at https://arxiv.org/abs/2306.09821 (2023).
38. Li, M. et al. Self-checker: plug-and-play modules for fact-checking with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024* 163–181 (Association for Computational Linguistics, 2024).

Acknowledgements

This study was funded in part by the National Medical Research Council of Singapore (HPHSR CLINICIAN SCIENTIST AWARD, NMRC/MOH/HCSAINV21nov-0001). National Key R & D Program of China (2022YFC2502800),National Natural Science Fund of China (82388101, 82103908, 82000417), Beijing Natural Science Foundation (IS23096), the Shandong Provincial Natural Science Foundation (ZR2021QH014), Shuimu Scholar Program of Tsinghua University (2023SM196), National Postdoctoral Innovative Talent Support Program (BX20230189), the China Postdoctoral Science Foundation (2024M751733) and the Fundamental Research Funds for the Central Universities. The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Author contributions

Conceptualization: H.J. and Y.C.T.; Methodology: H.J. and Y.C.T.; Investigation: H.J., X.W., C.H.S., J.Y., S.T.L., A.H.D., Y.C., N.Z., M.G., F.L., Z.W.L., Y.X.W., B.S., T.Y.W., S.C., K.K.Y., and Y.C.T.; Supervision: Y.C.T., S.C., and Y.K.K.; Writing – original draft: H.J. and Y.C.T.; Writing – review and editing: H.J., X.W., C.H.S., J.Y., S.T.L., A.H.D., Y.C., N.Z., M.G., F.L., Z.W.L., Y.X.W., B.S., T.Y.W., S.C., K.K.Y., and Y.C.T.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43856-025-00802-0.

Correspondence and requests for materials should be addressed to Yih-Chung Tham.

Peer review information *Communications Medicine* thanks the anonymous reviewers for their contribution to the peer review of this work.

[1]Beijing Visual Science and Translational Eye Research Institute (BERI), Eye Center of Beijing Tsinghua Changgung Hospital, School of Clinical Medicine, Tsinghua Medicine, Tsinghua University, Beijing, China. [2]Department of Cardiology, The Affiliated Hospital of Qingdao University, Shandong, China. [3]Beijing Key Laboratory of Intelligent Diagnostic Technology and Devices for Major Blinding Eye Diseases, Tsinghua Medicine, Tsinghua University, Beijing, China. [4]Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing, China. [5]Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [6]Department of Cardiology, National University Heart Centre Singapore, Singapore, Singapore. [7]Department of Cardiology, National Heart Centre Singapore, Singapore, Singapore. [8]Duke-NUS Medical School, Singapore, Singapore. [9]Division of Cardiology, Department of Medicine and Clinical Science, Yamaguchi University Graduate School of Medicine, Yamaguchi, Japan. [10]Dean's Office, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [11]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [12]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. [13]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. [14]Department of Cardiology, Smidt Heart Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [15]Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [16]Ophthalmology and Visual Science Academic Clinical Program, Duke-NUS Medical School, Singapore, Singapore. [17]These authors contributed equally: Hongwei Ji, Xiaofei Wang. [18]These authors jointly supervised this work: Susan Cheng, Khung Keong Yeo, Yih-Chung Tham. ✉e-mail: thamyc@nus.edu.sg