Paradigm Academic Press
Journal of Innovations in Medical Research
ISSN 2788-7022
APR. 2025 VOL.4, NO.2

Comparative Study on Response Efficacy of Generative Artificial Intelligence Large Language Model for Elderly Diabetes Mellitus

Ainingkun Xiang[1], Jingxue Tian[1], Dehua Hu[1] & Haixia Liu[1]

[1] College of Life Sciences, Central South University, Changsha, Hunan, China

Correspondence: Haixia Liu, College of Life Sciences, Central South University, Changsha, Hunan, China.

doi:10.63593/JIMR.2788-7022.2025.04.008

Abstract

We aimed to evaluate the response accuracy of different generative artificial intelligence (GAI) large language models to common problems of elderly diabetes, so as to compare the performance differences of various AI large language models in the quality of medical information service.

A standardized evaluation question pool containing 10 elderly diabetes related questions was constructed, and then four GAI chat robots using different generative artificial intelligence large language model were selected to answer the questions and score the accuracy of all answers. In addition, the problem is summarized into two dimensions of “diagnosis and evaluation” and “control and treatment”, and the above four GAI big language models are analyzed in these two dimensions.

In general, Moonshot model and Lark model are significantly better than DeepSeek LLM and SparkDesk model in response to common problems of elderly diabetes, with higher accuracy and strong stability, but there is no significant difference in response performance between Moonshot model and Lark model. In addition, in the dimensions of “diagnosis and evaluation” and “control and treatment”, Moonshot model and Lark model have better performance than DeepSeek LLM model and SparkDesk model.

Keywords: generative artificial intelligence, chat robot, large language model, senile diabetes mellitus, medical informatics

1. Introduction

Generative artificial intelligence (GAI) is an important branch of artificial intelligence. It is a technology that generates text, pictures, sounds, videos, codes and other content based on Algorithms and models.

At present, the AI technology system presents a diversified development trend, and its application scenarios have covered all fields of social production and life, which has attracted people’s attention. It is worth noting that there are significant differences in the technical complexity and intelligence level of different AI systems, and this heterogeneity directly affects its application effect and promotion value.

At present, AI has been widely used in the medical field. Many systems such as “AI triage” and “AI seeking medical treatment” have emerged. The acceptance and trust of medical staff and patient groups in AI driven medical information retrieval and analysis services have significantly increased. AI plays an active role in medical service efficiency, health management methods, etc., but it also faces risks and challenges in data privacy and security, algorithm deviation, ethics, etc. (Li et al., 2025). Therefore, the research on the application of artificial intelligence in medical treatment has a strong practical value.

In 2025, China’s generative AI model enters a new stage of development. DeepSeek, a large language model of artificial intelligence independently developed by China, has three major characteristics of “easy to use, open source, and free”. It has caused significant repercussions in the AI field, has been favored by many systems and

users, and has triggered huge discussions around the world. In terms of data, only 20 days after the DeepSeek application went online, its daily active users reached 22.15 million. At the same time, there have also been many related studies, such as taking DeepSeek as an example, discussing the technological transition, institutional synergy and technological civilization reconstruction in the digital Paradigm Innovation in the post ChatGPT era (Ling, 2025); Starting from DeepSeek, this paper discusses the supervision of generative artificial intelligence (Deng et al., 2025). At the same time, there are other AI with high popularity at present, such as the Doubao model, which has stood out in the fierce competition and achieved the counter attack of last mover first mover (Lei, 2025). The iFLYTEK spark has jointly developed the industry model with more than 20 industry enterprises, and the load of iFLYTEK spark app has exceeded 100million times (Liu, 2024). Kimi, launched in October 2023 by the dark side of the moon, an AI start-up, is the world's first intelligent assistant product that supports the input of 200000 Chinese characters (Zhao, 2024). This study selected four large-scale language models (Moonshot model, Lark model, DeepSeek LLM and SparkDesk model) as the research object, through the in-depth analysis and comparison of these cutting-edge intelligent dialog systems, this study can provide valuable references for academic research in related fields and benefit a wider range of user groups.

Senile diabetes refers to the metabolic syndrome caused by abnormal blood glucose metabolism in older people older than 60 years old. It is a common disease and frequently occurring disease, and has been widely concerned by older people. According to the data of the International Diabetes Federation, the number of diabetic patients aged 65 years and older in China is about 35.5 million. As the aging of the population continues to deepen, the number of elderly patients with diabetes is still on the rise. With the development of the Internet, more and more people use the Internet for elderly diabetes counseling. However, due to the problems of virtual and real information and information overload, the current situation of elderly diabetic patients obtaining health information from the network is not optimistic (Feng et al., 2024). Therefore, this study has important theoretical and practical significance for improving the quality of health knowledge acquisition of the elderly group, helping to improve the health literacy of the population.

In this study, the research team constructed a standardized assessment question pool containing 10 elderly diabetes related issues, and then selected four large-scale language models (Moonshot model, Lark model, DeepSeek LLM and SparkDesk model) as the research object, using its supported AI products (Kimi, Doubao AI DeepSeek, iFLYTEK spark AI) answered the questions and scored the accuracy of all answers. In addition, the questions were summarized into two dimensions of "diagnosis and evaluation" and "control and treatment", and the above four generative AI large language models were analyzed in these two aspects. The purpose of this study is to evaluate the response accuracy of different generative AI large-scale language models for common problems of elderly diabetes, so as to compare the performance differences of different generative AI large-scale language models in the quality of medical information service.

2. Materials and Methods

2.1 Design of Diabetes Related Issues

For older diabetes, the research team first designed a series of older diabetes related problem pool with high clinical value and public attention through literature review and expert consultation. Then 10 questions with clear answers in Guideline for the Management of Diabetes Mellitus in the Elderly in China (2024 edition) were selected, which covered key areas such as blood glucose control, complication prevention, lifestyle intervention, and the reference answers were given by referring to the standard. In addition, the problems are summarized into two dimensions of "diagnosis and evaluation" and "control and treatment". Subsequently, the research team selected four large-scale language models (Moonshot model, Lark model, DeepSeek LLM and SparkDesk model) as the research object, using its supported GAI products (Kimi, Doubao AI, DeepSeek and iFLYTEK spark AI) to test, input questions to each GAI under the same environmental conditions, and fully record its text output results.

In addition, before asking questions, the research team first input "Hello, next, I have a few questions to ask you, please give accurate and detailed answers as far as possible. Please try to answer according to the Chinese guidelines for the diagnosis and treatment of diabetes in the elderly (2024 version)", to ensure the consistency of the reference standards.

Table 1 shows 10 elderly diabetes related problems and dimension division.

Table 1. Details and dimensions of 10 questions

Question number	Question details	Dimension
1	What are the different types of senile diabetes	Null
2	Diagnostic criteria of diabetes mellitus in the elderly in China	Diagnosis

		and evaluation
3	What are the preventive measures for diabetes in the elderly (it is best to give each measure of tertiary prevention)	Null
4	I am an elderly diabetic. How can I adjust my lifestyle	Control and treatment
5	I am an elderly diabetic patient, only complicated with hypertension, without impairment of activities of daily living and instrumental activities of daily living. According to China's comprehensive health assessment criteria for elderly diabetic patients, is my health status good	Diagnosis and evaluation
6	I am an elderly diabetic who is using drugs with a high risk of hypoglycemia. After being evaluated according to China's comprehensive health assessment criteria for elderly diabetic patients (2024), my health is in good condition. According to the blood glucose control target of elderly diabetic patients in China, how much should I control my glycated hemoglobin, fasting or pre meal blood glucose, and bedtime blood glucose respectively	Control and treatment
7	I am an elderly diabetic with atherosclerotic cardiovascular disease. According to China's diabetes management standards, how much should I control my systolic blood pressure	Control and treatment
8	According to China's diabetes management standards, the hypoglycemia of elderly diabetic patients receiving drug treatment is divided into three grades	Diagnosis and evaluation
9	I am an elderly diabetic who is taking metformin, but has renal failure (estimated glomerular filtration rate is 40ml / [min · (1.73m2)]. Can I continue taking metformin? If not, what other drugs can I use	Control and treatment
10	Perioperative management of elderly patients with diabetes mellitus	Control and treatment

2.2 Score the Answers of the Four Generative AI Large Language Models

According to the GAI answers, the research team used the above four AI large-scale language model chat robots to score the accuracy of all the answers according to the reference answers formulated above (formulated according to Guideline for the Management of Diabetes Mellitus in the Elderly in China (2024 edition)) and the comparison with each other. Each item was scored on a 1-5 scale (1 is completely inaccurate, 5 is completely accurate), and the full score for each GAI is 50.

The specific operation is that the research team opens a new dialogue and inputs to each GAI under the same environmental conditions: "Hello, now please rate the accuracy of the following four answers A, B, C and D. The scoring standard is the correct answers I provide you below and the comparison between them. Please use a 1-5 score system for scoring (1 is completely inaccurate, 5 is completely accurate)." The format of the request for rating is: "the correct answer is: XX, the answer of a is: XX, the answer of B is: XX, and the answer of D is: XX."

2.3 Statistical Analysis of Score Results

According to the comprehensive situation, descriptive statistical analysis (Stata 18) was carried out on the 10 item scores of the four GAI, and the accuracy scores obtained and the stability of the scores were analyzed. Then, the normality test (Stata 18) was carried out on the overall data and the 10 item average scores of each AI, and Friedman test and pairwise comparison (SPSS 26) were carried out, and finally the statistically significant comprehensive ranking was obtained.

For the two dimensions of "diagnosis and evaluation" and "control and treatment", descriptive statistical analysis (Stata 18) was carried out on the score of the four GAI related dimensions, Shapiro Wilk W test normal test (Stata 18) was carried out, and Friedman test was used to analyze the significance of the difference (SPSS 26), as well as pairwise comparison after Bonferroni correction (SPSS 26), and finally a statistically significant comprehensive ranking was obtained.

3. Results

3.1 Raw Score Data of Four GAI

Table 2 shows the scoring and being scored of four AI large language model chat robots (iFLYTEK spark AI

(hereinafter referred to as XF), Doubao AI (hereinafter referred to as db), DeepSeek and Kimi) based on four generative AI large language models.

Table 2. GAI score data

Question number	GAI	Scores by GAI			
		xf	db	DeepSeek	Kimi
1	xf	2	3	3	3
	db	5	4	4	4
	DeepSeek	3	4	5	2
	Kimi	4	4	4	5
2	xf	3	3	4	3
	db	4	4	5	4
	DeepSeek	5	4	4	2
	Kimi	4	4	5	5
3	xf	3	3	3	3
	db	5	4	5	5
	DeepSeek	4	4	4	5
	Kimi	4	4	5	5
4	xf	3	2	3	2
	db	4	3	4	3
	DeepSeek	2	4	5	4
	Kimi	5	4	5	5
5	xf	4	4	4	3
	db	5	4	5	5
	DeepSeek	3	3	3	2
	Kimi	5	4	4	5
6	xf	4	2	4	3
	db	5	5	5	5
	DeepSeek	3	3	3	2
	Kimi	5	5	5	5
7	xf	3	3	3	3
	db	5	4	5	5
	DeepSeek	2	2	4	2
	Kimi	4	5	4	4
8	xf	2	1	3	3
	db	4	4	5	5
	DeepSeek	3	4	4	3
	Kimi	5	5	5	5
9	xf	5	4	4	3
	db	5	4	5	4
	DeepSeek	2	2	3	5
	Kimi	4	4	4	4
10	xf	3	2	3	3
	db	5	4	4	4

DeepSeek	4	2	5	4
Kimi	4	4	4	5

Notes: xf: iFLYTEK spark AI, db: Doubao AI, the same below.

3.2 Statistical Results of Four GAI

3.2.1 Descriptive Statistics

Descriptive statistical analysis was carried out on the scores of the ten questions of the four GAI (full score is 50), and the results are shown in Table 3.

Table 3. Descriptive statistical results of GAI score data

GAI	Average	Median	Total	Standard deviation
db	4.45	4.5	45	0.422
Kimi	4.50	4.5	45	0.333
DeepSeek	3.35	3.5	34	0.568
xf	3.05	3.0	31	0.538

(1) Accuracy: the scores of Kimi and Doubao are significantly higher than those of other GAI (the average scores are 4.50 and 4.45, respectively, with a median of 4.5), and the total scores are 45 and 45, respectively, indicating that they perform best in the accuracy of answers.

(2) The stability of accuracy score: Kimi's standard deviation is the smallest (0.333), the score fluctuation is the smallest, and the stability is the best; Doubao followed (standard deviation 0.422), while DeepSeek and iFLYTEK were less stable (standard deviation > 0.5).

3.2.2 Normality Test

For the overall data, Shapiro Wilk test showed that all 160 original scores did not meet the normal distribution ($w=0.979$, $p=0.018 < 0.05$). For the 10 item average score of each GAI, the average scores of the four GAI were in line with the normal distribution (P values > 0.05) after respective tests. Table 4 and Figure 1 show the results.

Table 4. Overall data and score test results of each GAI

Variable	Obs	W	V	z	Prob>z
xf	10	0.97644	0.363	-1.583	0.94332
db	10	0.93547	0.994	-0.009	0.50378
DeepSeek	10	0.9409	0.911	-0.159	0.5631
Kimi	10	0.95751	0.655	-0.697	0.75714
Overall	160	0.97948	2.523	2.105	0.01763

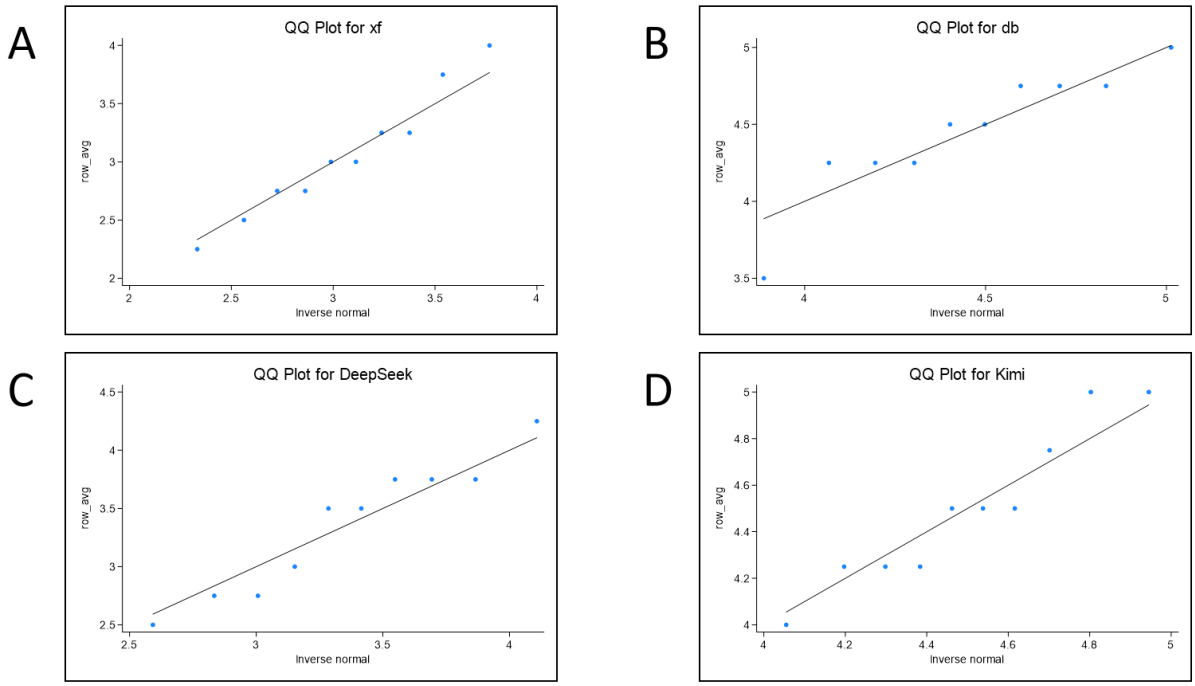

Figure 1. QQ plot for score test of four GAI

Notes: Score test QQ plot of four GAI (A) QQ plot of iFLYTEK spark AI score data; (B) QQ plot of Doubao AI score data; (C) QQ plot of DeepSeek score data; (D) QQ plot of Kimi score data.

3.2.3 Friedman Test and Pairwise Comparison

(1) Friedman test: the results showed that the scores of different GAI were significantly different ($\chi^2 = 29.4$, df=3, p<0.001).

(2) Pairwise comparisons after Bonferroni correction are shown in Table 5. The P value was compared with the corrected significance level (α =0.0083).

$$\alpha_{corrected} = \frac{0.05}{\text{Number of comparisons}} = \frac{0.05}{6} \approx 0.0083$$

Table 5. Statistical analysis results of GAI pairwise comparison

Comparison object	P value	significant（α=0.0083）
xf vs db	0.0001	Ture
xf vs DeepSeek	0.0012	Ture
xf vs Kimi	0.0001	Ture
db vs DeepSeek	0.0023	Ture
db vs Kimi	0.87	False
DeepSeek vs Kimi	0.0015	Ture

Significant difference group: Doubao AI vs iFLYTEK spark AI (p=0.0001), Doubao AI vs DeepSeek (p=0.0023), Kimi vs iFLYTEK spark AI (p=0.0001), Kimi vs DeepSeek (p=0.0015), DeepSeek vs iFLYTEK spark AI (p=0.0012).

There was no significant difference between Doubao AI and Kimi (p=0.87).

In summary, through Friedman test and Bonferroni correction, it was confirmed that Doubao and Kimi were significantly better than other GAI (p<0.0083).

3.2.4 Comprehensive Ranking

According to the comprehensive performance of accuracy (average score, total score) and stability (standard deviation), the ranking is as follows:

No.1 Kimi: The accuracy was the highest (average score is 4.50, total score is 45.0), and the stability of accuracy score was the best (standard deviation is 0.3333).

No.2 Doubao AI: The accuracy is slightly inferior to Kimi (average score is 4.45), and the stability of accuracy score is suboptimal (standard deviation is 0.4216), but there is no significant difference with Kimi.

No.3 DeepSeek: The accuracy was low (mean score is 3.35), and the stability of accuracy score was poor (standard deviation is 0.5676).

No.4 iFLYTEK spark AI: The accuracy is the lowest (average score is 3.05), and the stability of accuracy score is the worst (standard deviation is 0.5375).

3.3 Statistical Results of "Diagnosis and Evaluation", "Control and Treatment" Dimension

3.3.1 Descriptive Statistics of Two Dimensions

Table 6. Descriptive statistical analysis results of "diagnosis and evaluation", "control and treatment" dimensions

Dimension	GAI	Average	Median	Total	Standard deviation
Diagnosis and evaluation	DeepSeek	3.33	3.0	40	0.888
	Kimi	4.67	5.0	56	0.492
	xf	3.08	3.0	37	0.900
	db	4.50	4.5	54	0.522
Control and treatment	xf	3.10	3.0	62	0.788
	db	4.40	4.5	88	0.681
	DeepSeek	3.15	3.0	63	1.137
	Kimi	4.45	4.0	89	0.510

"Diagnosis and evaluation" dimension:

(1) Kimi's average score was the highest (4.67), indicating that the accuracy of his answer was the best; Doubao AI was the second (4.50), and its performance was also relatively excellent; The average scores of DeepSeek and iFLYTEK spark AI are low (3.33 and 3.08, respectively), and their performance is relatively poor.

(2) The median of Kimi and Doubao AI were 5 and 4.5, respectively, indicating that their score distribution was biased towards high scores; The median score of DeepSeek and iFLYTEK spark AI is 3, indicating that their score distribution tends to be medium or low.

(3) Kimi's total score was the highest (56), followed by Doubao AI (54), which was significantly better than DeepSeek (40) and Xunfei spark AI (37).

(4) Kimi's standard deviation was the smallest (0.492), indicating that its score fluctuation was the smallest and its stability was the best; The Doubao AI was the second (0.522), and its stability was good; The standard deviations of DeepSeek and iFLYTEK spark AI are relatively large (0.888 and 0.900, respectively), indicating that their scores fluctuate greatly and have poor stability.

"Control and treatment" dimension:

(1) Kimi's average score was the highest (4.45), indicating that the accuracy of his answer was the best; The Doubao AI was the second (4.4), and its performance was also relatively excellent; DeepSeek and iFLYTEK spark AI have low average scores (3.15 and 3.1, respectively) and relatively poor performance.

(2) The median of Doubao and Kimi were 4.5 and 4, respectively, indicating that their score distribution was biased towards high scores; The median score of DeepSeek and iFLYTEK spark AI is 3, indicating that their score distribution is biased towards medium or low scores.

(3) Kimi's total score was the highest (89), followed by Doubao AI (88), which performed significantly better than DeepSeek (63) and Xunfei spark AI (62).

(4) Kimi's standard deviation is the smallest (0.510), indicating that its score fluctuation is the smallest and its stability is the best; Doubao AI was the second (0.681), and its stability was good; The standard deviations of DeepSeek and iFLYTEK spark AI are relatively large (0.788 and 1.137, respectively), indicating that their scores fluctuate greatly and have poor stability.

3.3.2 Normality Test of Two Dimensions

The results of analyzing the normality of each GAI separately are shown in Figure 2.

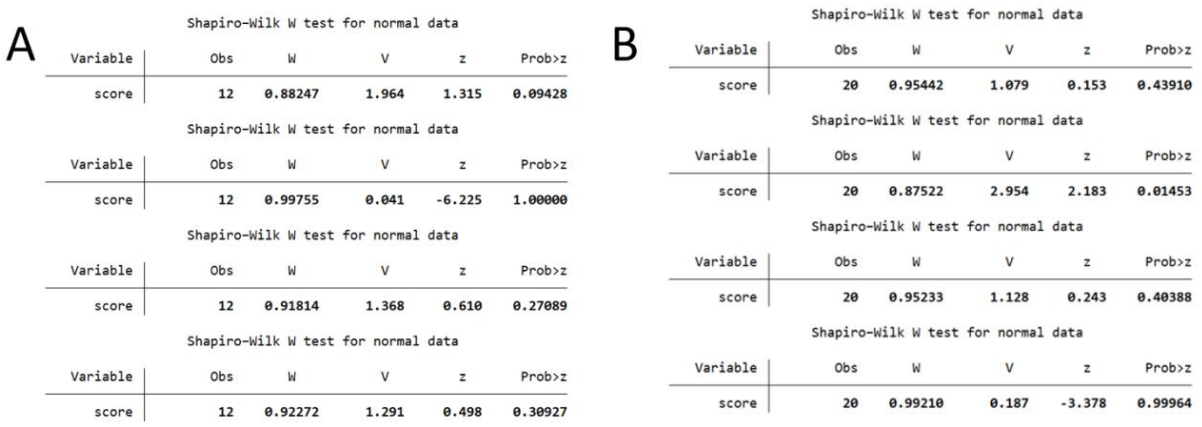

Figure 2. Normality test results of each GAI (A) Normality test results of "diagnosis and evaluation" dimension; (B) normality test results of "control and treatment" dimension

Notes: From top to bottom: iFLYTEK spark AI, Doubao AI, DeepSeek, Kimi.

"Diagnosis and evaluation" dimension: In the Shapiro Wilk W test normal test, the P values of the four GAI are > 0.05, and the original hypothesis is accepted, so the score distribution of the four GAI meets the normal distribution at the 0.05 significance level; Among them, the P value corresponding to Doubao AI is as high as 1.00, indicating that it is very close to normal and has good stability; In contrast, the P value corresponding to iFLYTEK spark AI is only 0.09428, indicating that its distribution may be skewed or abnormal, with poor stability.

"Control and treatment" dimension: In Shapiro Wilk W test normal test, iFLYTEK spark AI, DeepSeek and Kimi have P values > 0.05, and accept the original hypothesis, so the score distribution of these three GAI meets the normal distribution at the 0.05 significance level; The P value corresponding to Kimi is as high as 0.99964, indicating that it is very close to normal and has good stability; However, the P value of AI in Doubao =0.01453 < 0.05 rejected the original hypothesis, so its distribution did not conform to the normal distribution, indicating that its distribution may be skewed or abnormal, with poor stability.

3.3.3 Significance of Differences Between Two Dimensions by Friedman Test

Table 7. Friedman test results

Inspection statistics[a]		
Dimension		
	Diagnosis and evaluation	Control and treatment
Number of cases	12	20
Chi-square	21.471	25.575
Free degree	3	3
Asymptotic significance	0.000	0.000

Notes: a. Friedman test.

"Diagnosis and evaluation" dimension: Chi square value is 21.471, P value < 0.001, rejecting the original hypothesis, so there are significant differences in different GAI scores.

"Control and treatment" dimension: Chi square value is 25.575, P value < 0.001, rejecting the original hypothesis, so there are significant differences in different GAI scores.

3.3.4 Pairwise Comparison of Two Dimensions After Bonferroni Correction (Corrected α =0.0083)

Table 8. GAI pairwise comparison results

Inspection statistics[a]

		db - xf	DeepSeek - xf	Kimi - xf	DeepSeek - db	Kimi - DeepSeek	Kimi - db
Diagnosis and evaluation	Z	-3.002[b]	-0.540[b]	-2.840[b]	-2.547[c]	-2.676[b]	-1.000[b]
	Asymptotic significance (two tailed)	0.003	0.589	0.005	0.011	0.007	0.317
Control and treatment	Z	-3.841[b]	-0.354[b]	-3.582[b]	-3.065[c]	-3.265[b]	-0.258[b]
	Asymptotic significance (two tailed)	0.000	0.724	0.000	0.002	0.001	0.796

Notes: A. Wilcoxon signed rank test, B. based on negative rank, C. based on positive rank.

“Diagnosis and evaluation” dimension:

Comparing the P value in the pairwise comparison results with 0.0083, the P values corresponding to Doubao and iFLYTEK spark AI, Kimi and iFLYTEK spark AI, Kimi and DeepSeek are all less than 0.0083, so there is a significant difference between them, while the P values corresponding to DeepSeek and iFLYTEK spark AI, DeepSeek and Doubao AI, Kimi and Doubao AI are all bigger than 0.0083, so there is no significant difference between them.

“Control and treatment” dimension:

Comparing the P value in the pairwise comparison results with 0.0083, the P values corresponding to Doubao AI and iFLYTEK spark AI, Kimi and iFLYTEK spark AI, DeepSeek and Doubao AI, Kimi and DeepSeek are all less than 0.0083, so there is a significant difference between them, while the P values corresponding to DeepSeek and iFLYTEK spark AI, Kimi and Doubao AI are all bigger than 0.0083, so there is no significant difference between them.

3.3.5 Ranking of Two Dimensions

“Diagnosis and evaluation” dimension:

Score of this dimension: Kimi > Doubao AI > DeepSeek > iFLYTEK spark AI

Score stability of this dimension: Kimi > Doubao AI > DeepSeek > iFLYTEK spark AI

“Control and treatment” dimension:

Score of this dimension: Kimi > Doubao AI > DeepSeek > iFLYTEK spark AI

Score stability of this dimension: Kimi > Doubao AI > DeepSeek > iFLYTEK spark AI

Kimi has no significant difference with Doubao AI, but it is significantly better than DeepSeek and iFLYTEK spark AI (there is also no significant difference between them).

4. Discussion

This study shows that, in general, the Moonshot model (represented by Kimi) and the Lark model (represented by Doubao AI) perform best in answering questions related to older diabetes, with high accuracy and stability, and are recommended to be used preferentially in medical consultation. The performance of SparkDesk model (represented by iFLYTEK spark AI) and DeepSeek LLM (represented by DeepSeek) is relatively weak, and the answer logic needs to be optimized. In addition, in the dimensions of “diagnosis and evaluation” and “control and treatment”, the Moonshot model has better performance than the Lark model and the DeepSeek LLM model. The analysis of this study is based on strict statistical tests, ensuring the scientificity and reliability of the conclusions.

The reason for this phenomenon is the difference between the algorithm of generative AI and knowledge updating. Therefore, this study also provides a foundation for the in-depth study of generative AI. At present, there are also many researches on the algorithm of GAI, such as the discussion on the innovation and optimization of DeepSeek series models in large model training (Zhang, 2025), and for six serum tumor markers, eight different joint detection models are built in the modeling cohort and test cohort by combining eight different AI algorithms, and the joint detection model with the best performance is selected (Ren et al., 2025). This research is beneficial to the algorithm research and knowledge updating research of GAI.

At present, with the rapid development of artificial intelligence technology, its deep integration with the medical

and health field has become an important frontier direction of current research. As for the relevant research on the performance comparison of different AI in medicine, the current number of studies is relatively small, and the scoring method is human doctors' evaluation. Foreign researchers put forward ten common anesthesia questions to three AI chat robots: chatGPT4 (openAI), Bard (Google) and Bing chat (Microsoft). Five resident program directors from 15 medical institutions in the United States evaluated the answers of each chat robot in a randomized, blinded order (NGUYEN et al., 2024). Domestic researchers have also studied the differences between a variety of large-scale language models and the answers of ophthalmologists (Hu, 2023). Compared with them, this study formulated the reference answer according to the comprehensive evaluation standard of health status of elderly diabetic patients in China (2024), and used AI to score based on the reference answer and the comparison, which greatly reduced the influence of subjectivity. In addition, this study summarized the problems into two dimensions, "diagnosis and evaluation" and "control and treatment", and compared the two dimensions to enhance the depth of the study. In addition, the research object of this study is four AI large-scale language model chat robots (iFLYTEK spark AI, Doubao AI, DeepSeek, Kimi) with high popularity in China, which has greater significance for the current Chinese people's choice of AI.

The reliability of the application of artificial intelligence in the medical field has also received close attention. Current research focuses on whether artificial intelligence is reliable for disease detection. Some studies have evaluated the clinical safety of AI supported screen reading scheme compared with standard screen reading after mammography by radiologists, and found that compared with standard double reading, AI supported mammography screening produced similar cancer detection rate and greatly reduced screen reading workload, indicating that AI is safe to use in mammography screening (LåNG et al., 2023). Some researchers conducted a cluster randomized cross-over controlled trial to evaluate the impact of artificial intelligence — based diagnostic support software on the detection of proximal caries on wing X-rays, and proposed that AI could improve the diagnostic accuracy of dentists (MERTENS et al., 2021). As well, a randomized comparative effectiveness trial showed that AI-cbt-cp (cognitive behavioral therapy for chronic pain using artificial intelligence) was not inferior to the telephone CBT-CP (cognitive behavioral therapy for chronic pain) provided by the therapist, and the time needed by the therapist was greatly reduced (PIETTE et al., 2022). And, according to the Chinese and English nursing suggestions given by ChatGPT, some researchers evaluated its application value in chronic disease nursing (Yin et al., 2024).

In addition, this study also triggered the research team's thinking on the relationship between AI and traditional medical industry. AI technology is reshaping the medical industry, and there are questions like "Will the application of AI in the medical field replace doctors in the future?" (Zhu et al., 2025). The research team believes that the appropriate application of AI can promote the development of the medical industry. However, AI still faces many challenges in medical practice, including ethical dilemmas, data privacy issues and the lack of humanistic care. These limitations indicate that AI cannot completely replace the role of traditional doctors. The two promote each other and advance hand in hand is the ultimate solution of "AI+ medical" in the future.

There is also room for improvement in this study, such as evaluating the response performance of the generative AI large language model only from the dimension of "accuracy", and exploring the theme of "older diabetes". In addition, the evaluation system of the generative AI large language model is not absolutely scientific and accurate. More research dimensions, research topics, and research numbers will help enhance the value of research.

Author Contribution Statement

Ainingkun Xiang is responsible for the design of questions, the collection of scores and the writing of the first draft. Jingxue Tian is responsible for the statistical analysis of the original score data, and participated in the writing and revision of the article. Haixia Liu and Dehua Hu provided important guidance for this study in terms of topic selection, implementation, article revision, and detail optimization.

Declaration of Interest

All authors declare that there are no conflicts of interest.

Acknowledgements

This work was supported by National Social Science Foundation of China (Project No.: 20BTQ081); Research and development project in key fields of Hunan Province (Project No.: 2021WK2003).

References

DENG J P, ZHAO Y S., (2025). The breaking and changing situation of DeepSeek: on the regulatory direction of generative AI. *Journal of Xinjiang Normal University (Edition of Philosophy and Social Sciences)*, (04), 1-10.

FENG C Q, MENG L X, LUO G Q, et al., (2024). Research progress on the acquisition behavior of network health

information of elderly diabetic patients. *Chinese Medical Sciences*, *14*(21), 40-3.

HU C L., (2023). The application of the large language models in ophthalmic consultation. Guangdong: Shantou University.

LåNG K, JOSEFSSON V, LARSSON A M, et al., (2023). Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study. *Lancet Oncol*, *24*(8), 936-44.

LEI C., (2024). ByteDance Doubao is very popular, and the AI industry chain is collectively restless. *21st Century Business Herald*, 012.

LI T, ZHANG J, LI Y Z, et al., (2025). Innovative Applications, Risk Challenges and Governance Countermeasures of Artificial Intelligence in the Field of Healthcare. *Journal of Medical Informatics*, *46*(01), 2-8+16.

LING X X., (2025). DeepSeek Ushers in the Post-ChatGPT Era: On Digital Paradigm Innovation and Its Operational Philosophy. *Journal of Northwestern Polytechnical University (Social Sciences)*, 1-9.

LIU Q F., (2024). The latest development and industrial application of Spark model technology. *China Economic Report*, (04), 130-3.

MERTENS S, KROIS J, CANTU A G, et al., (2021). Artificial intelligence for caries detection: Randomized trial. *J Dent*, *115*, 103849.

NGUYEN T P, CARVALHO B, SUKHDEO H, et al., (2024). Comparison of artificial intelligence large language model chatbots in answering frequently asked questions in anaesthesia. *BJA Open*, *10*, 100280.

PIETTE J D, NEWMAN S, KREIN S L, et al., (2022). 0 Patient-Centered Pain Care Using Artificial Intelligence and Mobile Health Tools: A Randomized Comparative Effectiveness Trial. *JAMA Intern Med*, *182*(9), 975-83.

REN N N, ZHANG H P, JIN S Y., (2025). Value of a combined detection model of six serum tumor markers and artificial intelligence algorithm in the diagnosis of lung cancer. *Zhejiang Medicine*, *47*(03), 268-73+339.

YIN B Q, LIU S Y, WANG H R, et al., (2024). Review on ChatGPT in Chronic Disease Nursing Care. *Military Nursing*, *41*(02), 83-5.

ZHANG H M., (2025). How DeepSeek-R1 was created? *Journal of Shenzhen University (Science and Engineering)*, 1-7.

ZHAO H., (2024). Domestic AI model Kimi's "debut" aims at the long text track. *China Strategic Emerging Industry*, (13), 72-5.

ZHU J L, ZHANG H, CHEN H., (2025). Will AI replace doctors? *Yangcheng Evening News*, A04.